Language Testing and Evaluation 48

David Wei Dai

Assessing Interactional Competence

Principles, Test Development and Validation through an L2 Chinese IC Test

PETER LANG

With the growing recognition of the need to broaden the definition of Interactional Competence (IC) for communication and learning, this monograph offers the first book-length treatment on the conceptualization, development and validation of IC assessment instruments. Combining psychometrics with discourse analysis, highlights of the book include: 1) evidence that a holistic IC construct – encompassing the sequential, emotional, logical, moral and categorial dimensions – can be assessed reliably, 2) a practical IC rubric that is adaptable to diverse languages and contexts, 3) demonstration that L2 speakers can have stronger IC than L1 speakers. The book argues that IC needs to be taught and assessed for both L1 and L2 speakers to promote fairness in language education.

In this superb, meticulously designed, intellectually coherent book based on award-winning scholarship, David Wei Dai takes the reader on a riveting journey tackling key challenges in assessing Interactional Competence. Ingenious and groundbreaking; there is no looking back.

(Talia Isaacs, University College London)

David Wei Dai's book is an exemplary study of test development and validation. It breaks new ground in the assessment of Interactional Competence, and is an invaluable resource for novices and seasoned researchers alike.

(Carsten Roever, University of Melbourne)

David Wei Dai PhD FHEA is Lecturer/Tenured Assistant Professor in Professional Communication at UCL Institute of Education, University College London. His research interests include language assessment, Interactional Competence, professional communication and discourse analysis. His work has appeared in international peer-reviewed journals including *Applied Linguistics, Language Teaching Research, Medical Education, Applied Linguistics Review, Language Assessment Quarterly, Language, Culture* and *Curriculum* and *Journal of English for Academic Purposes.*

Assessing Interactional Competence

Language Testing and Evaluation

Series editors: Claudia Harsch and Günther Sigott

Volume 48

Zur Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch die Herausgeber der Reihe geprüft.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is reviewed by the editors of the series.

David Wei Dai

Assessing Interactional Competence

Principles, Test Development and Validation through an L2 Chinese IC Test

PETER LANG

Berlin - Bruxelles - Chennai - Lausanne - New York - Oxford

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.dnb.de.

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

ISSN 1612-815X
ISBN 978-3-631-88250-4 (Print)
E-ISBN 978-3-631-88585-7 (E-Book)
E-ISBN 978-3-631-88586-4 (E-PUB)
DOI 10.3726/b21295

Published by: Peter Lang GmbH, Berlin, Deutschland

info@peterlang.com - www.peterlang.com

PETER LANG
Open

Dedication: To all those I've offended in interaction, mostly unintentionally, in my L2s as well as my L1s.

Preface

This book is an extension and adaptation of my PhD thesis – Dai (2022). The thesis has won the 2023 American Association for Applied Linguistics (AAAL) Dissertation Award (https://www.aaal.org/dissertation-award). This is the first time that an Australian PhD thesis has won the award.

The data collection and writing of this book were supported by the Australian Government Research Training Program Scholarship, the Faculty of Arts Graduate Research International Grant from the University of Melbourne, the Research Data Collection Support from the School of Languages and Linguistics at the University of Melbourne, the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment, and the British Council Assessment Research Award. This book is published Open Access with the generous support from University College London.

Some parts of the book have appeared in Dai (2023a) published in *Language Teaching Research* and Roever and Dai (2021) published in *Assessing speaking in context: Expanding the construct and its applications* by Multilingual Matters.

Due to the complexity of documenting a test development and validation project, I start each chapter with a brief summary that encapsulates the key messages of the chapter. References to sections, tables and figures in the text have hyperlinks embedded so readers can be directed to relevant parts of the book if/when needed. For readers that are especially pressed for time, they are encouraged to start with the **Summary of the Book**, **Chapter 1** Introduction and **Chapter 8** Conclusion to get an overview of the monograph. **Chapter 3** lays out the interpretive argument for validation, **Chapters 4–6** present three interrelated studies designed to gather the validity evidence motivated in **Chapter 3**, and **Chapter 7** evaluates the evidence garnered in **Chapters 4–6** to support validation in the form of a validity argument.

Foreward

I first became aware of David Wei Dai's work while he was studying for a Master of Applied Linguistics at the University of Melbourne, Australia. David and I happened to share a Masters dissertation supervisor – Carsten Roever – albeit around 12 years apart in time. David's dissertation (or, following the Melbourne University convention, his 'minor thesis') was on a topic that continues to hold a lot of interest for me: the effects of speaker accent in second language listening assessment. I was pleased, both for David and for my former supervisor, when I learned that David had won the Caroline Clapham IELTS Masters Award for his work (which has since been published in *Language Assessment Quarterly*). Since that time, I've been following David's academic journey as he turned to focus on another topic that presents conceptual and practical challenges for language testers: interactional competence. David continued to work with Carsten Roever (moving towards pragmatics and interaction was perhaps inevitable!), and this book represents the product of his PhD research, also conducted at the University of Melbourne, which focused on the assessment of interactional competence among second language (L2) learners of Chinese.

Interactional competence has been a vibrant area of research in language assessment for some time. Many researchers have drawn on the principles and methods of conversation analysis (CA) and membership categorization analysis (MCA) to understand how interaction is managed within speaking test performance, providing empirically based insights into test-taker discourse using a rigorous methodological approach. It is ironic, in some ways, that this line of research has grown at precisely the same time that many large-scale tests have explored automated delivery and scoring of speaking, with a corresponding shift away from the messiness of human-human communication towards more tightly controlled, psycholinguistic-oriented speaking tasks. As I have expressed elsewhere (e.g., Harding & Fulcher, 2022), we are at a critical juncture in the evolution of language assessment design, where our understanding about the nature of communicative competence is developing in greater depth, while the affordances of technology sometimes seem to be moving assessment practices in an altogether different direction.

Within this broader context, David's book is timely and important. It offers novelty in various respects. David focuses on L2 learners of Chinese, for instance, addressing an urgent need for more research on assessing IC in languages other than English. The setting of the study is a computer-mediated communication

(CMC) environment, which continues to gain prominence in the assessment landscape after the Covid era. David also employs a diverse range of methods to answer specific questions within an argument-based approach to validation. It is a strong and coherent piece of research, and I am not surprised that the thesis on which this book is based has won the 2023 American Association for Applied Linguistics (AAAL) Dissertation Award.

For me, though, what sets David's work apart as particularly useful for other scholars of IC is its theoretical robustness. David does not shy away from the fundamental conundrums at the heart of IC assessment, and he seeks to unify otherwise separate philosophical approaches to chart a way through for test developers. Most impressive of all is David's vision of IC, which moves beyond the categories familiar to many language testers and draws on a wide range of thought within sociology and philosophy to develop, in his words, a theory of IC that encompasses 'not only the sequential but also the emotional, moral, logical and categorial dimensions of interaction' (p. 12). This is an important contribution that is likely to have resonance with fields beyond language testing and assessment. It is a useful example of how thinking through applied problems can result in theoretical advances.

Readers will find this an erudite and thought-provoking book that will challenge their pre-conceptions about IC and how it might be assessed. That was certainly my experience. In fact, I was lucky enough to have an early glimpse of David's research and thinking when he visited Lancaster in early-September 2022. David presented his work to our Language Testing Research Group and fielded some very probing questions from Lancaster students. I found myself thinking about his new model of IC, and its implications, for some time after that visit. And I know that I will be referring to David's work frequently in the future.

I hope you enjoy engaging with this unique perspective on IC, and I look forward to seeing how David's work continues to raise new ideas and new possibilities for assessment.

Luke Harding
Lancaster
December, 2023

Summary of the Book

This book documents the development and validation of a standardized large-scale language test on second language (L2) interactional competence (IC), which reflexively informed the conceptualization of IC and theorization of social interaction as a competence. More specifically, this book investigates L2-Chinese as its target language for an IC test that is delivered in the computer-mediated environment to enhance the practicality of IC assessment. Throughout the process of test development and validation, I defined IC as a theoretically multidimensional construct, which encompasses a speaker's ability to manage interaction at the sequential, emotional, moral, logical and categorial dimensions. From the assessment perspective, I demonstrated that IC is concurrently theoretically multidimensional and psychometrically unidimensional, which corroborates the assessability of IC as a test construct.

Adopting an argument-based approach to test validation, I designed the test following a task-based needs analysis (TBNA), eliciting the perspectives from L2-Chinese speakers, first language (L1) Chinese teachers, and L1-Chinese interactants on what L2-Chinese speakers struggle the most with interpersonal interaction in L2 Chinese. Findings from the TBNA informed the design of a nine-item IC test that targets test-takers' ability to manage social actions that are disaffiliative in nature, which are actions that can threaten social harmony (e.g., making a complaint about workplace fairness to your employer). The test is delivered on a mobile-phone application, covers three sub-language use domains (everyday life, work, and study), and includes three degrees of interactiveness in terms of task methods (1st pair part voice messaging, 2nd pair part voice messaging, and live video chat).

The specification of the IC test construct and development of the rating scale were based on everyday-life linguistic laypersons' criteria of IC. The use of everyday members' criteria to define the test construct represents an attempt to democratize research by complementing the perspective of applied linguists with the one of linguistic laypersons, who use the target language on a daily basis and are the final arbiters of successful interaction. 36 linguistic laypersons listened to and commented on 22 pilot test-takers' performances on the test. Thematic analysis of linguistic laypersons' interview transcripts and written comments returned five indigenous categories that formed the five rating categories in an indigenous IC scale. When analysing pilot test-taker discourse in terms of IC, I proposed the use of Sequential-Categorical Analysis, which combines

Conversation Analysis and Membership Categorization Analysis to allow for the investigation of both the sequential and categorical aspects of interaction. The analysis of test-taker discourse through Sequential-Categorical Analysis assisted me in theorizing linguistic laypersons' a-theoretical indigenous rating scale into a theorized IC scale, which has (1) disaffiliation control, (2) affiliation promotion, (3) morality, (4) reasoning, and (5) social role management as its five rating categories. The IC construct developed in this book moves beyond constructs in existing IC scholarship as it illustrates the multidimensional nature of IC, encompassing not only the sequential dimension of interaction, but also the emotional, moral, logical and categorical dimensions.

105 test-takers from 26 different countries participated in the main testing study, whose performances were rated in a fully-crossed design by two raters using the theorized IC rating scale. Many Facet Rasch Analysis on rating results showed that item performance, rater reliability and rating scale functioning were satisfactory. Rasch Principal Component Analysis demonstrated the unidimensionality of IC as a test construct. Correlational analyses on test-takers' IC test performance and an external measure of proficiency revealed that proficiency had limited predictive strength on IC: $r\ (104) = .42$, $p<0.05$. Lower proficiency L2 speakers could outperform higher proficiency L2 speakers and even L1 speakers. This finding challenges the longstanding native-speakerism in language teaching and testing as it shows that L1 speakers are not the gold standard of IC. It also highlights that IC does not develop automatically as L2 speakers' proficiency increases. Therefore, IC is an ability that needs to be taught and assessed separately from proficiency for L1 and L2 speakers alike.

The extrapolative power of the IC test was ascertained through a self-assessment questionnaire and a peer-assessment questionnaire. Test-takers' self-assessment and peer-assessment were correlated with their test performance, the results of which showed that the IC test can reasonably predict test-takers' IC in non-assessment, real-life settings. Test-takers' attitude towards the test was also elicited in the questionnaire, which was favourable and supportive of the use of IC items in general speaking assessment.

Although the role-play IC test in this book is based on L2 Chinese, findings from this book are potentially applicable to other test tasks and target languages due to the highly theorized nature of the IC construct in this book, which is not specific to any context, language, or task type. The theorized IC rating scale embodies a holistic IC construct that goes beyond the mechanics of interaction and ventures into the assessment of speakers' ability at managing affect, logic, morality and categorization in and for interaction. This IC model is theoretically robust as it is consistent with other more holistic models on interpersonal interaction such

as Dell Hymes's original conceptualization of communicative competence and Aristotle's three artistic proofs. Future research can adapt the current IC test and localize the theorized IC scale to see if findings from this study still obtain when applied to other target languages, task types and assessment contexts. The use of computer-mediated communication platforms for test delivery in this book increases the practicality, accessibility and affordability of IC assessment. The computer-delivered nature of this IC test also allowed test-takers from a wide range of backgrounds to participate, which differed from the typical participants in applied linguistics research, who tend to be affluent middle-class university students. This design in the book helped to promote fairness, democracy and inclusivity in applied linguistics and language testing research.

中文概述 (Summary of the Book in Chinese)

本书记录了作者开发并验证一个互动能力（interactional competence, IC）考试的过程。作者首先在理论层面定义了互动能力具有多维特征，包含控制序列(sequential)，情感(emotional)，道德(moral)，理性(logical)和人物关系分类(categorial)能力。同时作者从测试的角度证明互动能力同时具有理论多维性（theoretically multidimensional）和测量单维性（psychometrically unidimensional）的特征，为测量互动能力提供了方向。

本研究所开发的互动能力考试是一个面向汉语二语考生设计的计算机辅助考试。依据Michael Kane的基于论证的效度检验框架，作者首先采取任务式教学的需求分析，了解汉语二语者在汉语二语人际交往中感觉最困难的地方。作者在需求分析中收集了三方不同的观点，包括汉语二语者，汉语母语语言老师，和经常与汉语二语者交流的汉语母语者。根据任务式教学需求分析的结果，作者开发了一项互动能力考试，共包含九个题目。该测试在手机软件上进行，主要测量考生处理各种非亲和性社会行为的能力，考题涵盖日常交流，职场沟通和学术对话三个语言使用领域。依据互动程度的高低，考题任务包括三种类型：相邻对第一部分（1st pair part）语音短信，相邻对第二部分（2nd pair part）语音短信和即时视频对话。

接下来，作者基于日常生活中非专业人士（linguistic laypersons）关于互动能力的标准，设计了测试构念（construct）说明和评分量表。 考虑到非专业人士日常使用语言并是语言能力的最终裁决方，使用他们的标准来定义测试构念能与语言学专业人士的视角形成互补，促进应用语言学研究民主化。作者收集了36位语言学非专业人士对参加预测的22位考生样本的评论笔记，并对这些非专业人士进行了访谈。通过对评论笔记和访谈文本的主题分析（thematic analysis）， 作者获取了5个本地标准（indigenous criteria），进而由此设计了一个包含五个评分维度的本地互动能力量表（indigenous IC scale）。

随后，作者从互动能力的角度对22位预测考生的话语数据进行定性分析，在此过程中作者定义并提出一种新的研究方法：序列分类分析法（Sequential-Categorial Analysis）。序列分类分析法结合了会话分析(Conversation Analysis)和社会成员分类分析(Membership Categorization Analysis)，从序列（sequential）和分类（categorial）双重角度来分析人际互动和人际关系。通过对考生话语数据进行序列分类分析，作者将非理论的（a-theoretical）本地互动能力量表进行理论化。该理论化的互动能力量表包括五个评分维度：1）控制冲突（disaffiliation control），2）拉

近关系（affiliation promotion），3）人品素质（morality），4）理性思辨（reasoning），5）身份意识（social role management）。以往文献对互动能力的定义多停留于时间序列层面(sequential)，而本研究中对互动能力测试构念的界定不仅基于序列，还 涉及到人际互动中情感(emotional)，道德(moral)，理性(logical)和人物关系分类(categorial)等层面。这凸显了互动能力在理论层面多维度（multidimensional）的特征。

接下来，来自26个国家的105位考生参加了正式的互动能力测试。作者聘请两位评分员使用理论化的互动能力量表分别对全部考生的表现进行打分。对评分结果的多面Rasch分析（Many facet Rasch analysis）结果显示测试题目维度（如难度、区分度），评分员信度和评分量表功效都较为理想。Rasch主成分分析证明了互动能力的测量单维性。对考生互动能力和语言水平（proficiency）的相关分析结果显示语言水平能力对互动能力的预测力很有限: $r(104) = .42$, $p<0.05$。研究发现语言水平较低的二语考生往往比语言水平较高的二语考生和一语考生展现出更强的互动能力，这说明母语使用者并不是互动能力的黄金标准，该结果对语言教学和语言测试中长期存在的母语中心主义提出了挑战。研究结果也证实互动能力并不会随着二语使用者语言水平的提高而提高。因此，对于互动能力的教授和测评需要与语言水平区分开来，且互动能力对于母语和二语使用者来说均是需要学习和考核的能力。

最后，作者使用了自我评测问卷和同行评测问卷对所开发的互动能力测试的外推力(extrapolative power)进行验证。作者将考生的自我评测得分和同行评测得分与互动能力测试得分进行相关分析，结果显示互动能力测试得分能在一定程度上外推考生在日常生活中的互动能力表现。考生自我评测问卷中也收集了考生对本互动能力测试的态度，结果显示考生对该测试表现出较为积极和支持的态度，希望今后的考试中能增加对互动能力的考核。

虽然本研究所开发的角色扮演互动能力测试基于二语中文，由于本研究对互动能力构念的定义具有高理论化的特征，本研究所采用的研究方法和研究结果对于其他测试任务类型，目标语言，测试场景也具有广泛的适用性。此外，作者提出的理论化的互动能力量表打破了原有对互动能力定义的局限，扩展到测量一个人如何通过控制情感、逻辑、道德和人物关系分类，从而实现更好地互动。本研究中对互动能力构念的全新界定与戴尔海姆斯（Dell Hymes）提出的交际能力概念（communicative competence）和亚里士多德对修辞策略（artistic proofs）的定义有着异曲同工之处，这展现出本研究中互动能力构念的理论稳健性。未来的研究可以尝试对本书所开发的互动能力测试进行适当调整，将理论化的互动能力量表本地化，进而考察本研究结果在其他目标语言，测试任务和测试场景中的推普性。最后，该测试采用计算机辅助技术的测试方法，使得该测试的实用性、可及性和可负担性得到了提升，特别在例如因受

到新冠疫情的影响，现场测试难以实施的情况下，这一测试方法的优势更加凸显。计算机辅助的特点也为各种不同背景的考生提供了参与到本研究中的机会，因此本研究招募的样本不同于其他应用语言学研究中的大学生样本，他们大多来自富裕中产阶层。从这个意义上讲，本研究在推进应用语言学领域中的公平性、民主性和包容性方面做出了一定的贡献。

Acknowledgements

Writing an acknowledgement for a research monograph like this is a truly humbling experience, not because completing the book itself is anything extraordinary, but because it has relied on the support and generosity of so many people.

First and foremost, I am grateful to Prof Carsten Roever, who provided invaluable guidance in the key areas where this book is positioned: language assessment, second language pragmatics, and interactional competence. I am indebted to Dr Helen Zhao, whose SLA perspective challenged, complemented, and enriched this assessment project. A special thank you to Prof Ute Knoch, whose meticulous feedback and attention to detail have improved the design and write-up of this book.

I owe much of what I have learnt about applied linguistics and language assessment to the scholars who have taught me in the School of Languages and Linguistics at the University of Melbourne. I thank the late Prof Tim McNamara, Dr Brett Baker, Dr Neomy Storch and Dr Paul Gruba for their encouragement, support, and guidance at various stages of my training.

I am deeply appreciative to Dr Xun Yan, Dr Peixin Zhang, Dr Jason Fan, Prof Yan Jin, Dr Ivy Chen, and Dr Naoki Ikeda for generously sharing with me their experience and wisdom in language assessment. I thank Dr Mike Linacre for patiently answering my queries on Rasch. I have benefited greatly from discussing with Dr Yang Zhang, Dr Alfred Rue Burch, and Dr Michael Davey on the conversation analytic side of this project. I am grateful for the support from Dr Daniel Lam, Dr Aung Si, and Dr Csaba Z Szabo when I was navigating the lengthy journey of writing this book.

When I started the book, I did not foresee a test development and validation project would involve so many groups of participants. The book would not exist if it were not for the goodwill and generosity of the many language enthusiasts, language educators and other stakeholders. I am especially indebted to Dr Jian Xu, Dr Sichang Gao, Dr Bo Hu, Dr Yunwen Su, Dr Alex Monceaux, Mr Eitan Waxman, Ms Yuxuan Chen, Mr Aleks Novakovic, Dr Jindan Ni, and Mr Kevin Yang, who were instrumental in assisting me with data collection. The collection of data would also not have been possible if it was not for the financial support from the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment, the British Council Assessment Research Award, and two data collection grants from the University of Melbourne.

A special and heartfelt thanks to my previous manager Dr Sheila Vance and colleagues at the Faculty of Medicine, Nursing and Health Sciences at Monash University. Dr Vance's unflagging support has ensured that I was able to complete the book in a timely manner.

I am grateful to Prof Luke Harding for penning the Foreword of the book, Dr Talia Isaacs and Prof Carsten Roever for the generous endorsements, and Prof Claudia Harsch and Prof Günther Sigott for the detailed and insightful feedback on the book manuscript.

Finally, I would like to acknowledge my parents, whose love, encouragement, and unwavering support have made it possible for me to be the person I am today. They have always been my bedrock during the most challenging times. They have instilled in me the notion of hard work, self-respect, and discipline being fundamental to realizing my full potential and achieving lifelong goals. I am grateful to Dr Soo for his wisdom and passion, which inspire me with hope and confidence whenever I doubt myself or lose motivation. Lastly, I would like to thank my dearest friends Sophie, Henry, Jonathan, Mikey and George for their sincerity, good humour and always being there for me.

Table of Contents

List of Tables

List of Figures

List of abbreviations

BMA:	broader membership apparatuses
CA:	conversation analysis
CMC:	computer-mediated communication
DCT:	discourse completion task
DE:	domain expert
EAP:	English for academic purposes
F2F:	face-to-face
F2FC:	face-to-face-communication
H-S interviews:	Hermeneutic-Socratic interviews
HSK:	Hanyu Shuiping Kaoshi (Chinese Standard Exam)
IC:	interactional competence
L1:	first language
L2:	second language
LC:	linguistic competence
MCA:	membership categorization analysis
NA:	needs analysis
ODA:	Official Development Assistance
OECD:	Organization for Economic Co-operation and Development
OET:	the Occupational English Test
PCA:	principal components analysis
SD:	standard deviation
SE:	standard error
TBNA:	task-based needs analysis
TLU domain:	target language use domain
UG:	universe of generalization

Chapter 1 Introduction

The assessment of second language (L2) speaking ability has long followed the psycholinguistic-individualist approach that focuses on a test-taker's ability to produce speech and assesses this ability based on linguistic components such as lexical range, grammatical accuracy, and pronunciation. This approach is adopted in most major language tests (e.g., IELTS and TOEFL) but is not without controversies. One of the main issues language assessment specialists takes with the psycholinguistic-individualist approach is that it under-represents the social, interactional reality of speaking (Roever & Kasper, 2018). Speaking in real life most frequently takes the form of talking, which is interaction with other people. If this key feature is underrepresented in speaking tests, scores from the speaking tests cannot provide reliable information to inform end-users about how test-takers actually perform in real life, threatening the validity of the speaking tests and the legitimacy of the testing practice.

The assessment of interaction, or more specifically, a speaker's interactional competence (IC), questions the current practice of speaking assessment and concurrently raises new challenges. The conceptualization of IC draws heavily from the field of Conversation Analysis (CA) on everyday interaction. However, CA is empiricist in its analytic approach and does not evoke etic standards in its depiction of interaction. This is misaligned with the practice of assessment in which test-taker performances need to be benchmarked against standards. This philosophical tension between CA and testing has long troubled IC assessment researchers and needs to be acknowledged and addressed before IC assessment can proceed.

In terms of the current scope of IC assessment research, previous studies have largely focused on L2-English as a target language and face-to-face (F2F) communication as a medium of assessment (Ikeda, 2021; Youn, 2015). More research is needed in other languages to enrich our understanding of IC assessment. The F2F mode of test delivery, despite its authenticity, highlights the logistic and practicality challenges of IC assessment as it can be expensive to create F2F interactional settings for test-takers to interact with interlocutors in order to assess their IC. The COVID-19 global pandemic further prompts test developers to rethink the practice of language assessment and explore alternative platforms for test delivery that do not require F2F interaction. More research is needed to see if IC assessment can be conducted in the online space to lower the

cost and make language assessment operationalizable when F2F assessment is not feasible.

Since IC assessment is still relatively new in the field of L2 speaking assessment, IC as a test construct is under-specified and there is no consensus on how IC should be best defined. Previous research on IC development and assessment has focused on various indicators of IC, but the choice of the IC indicators is either based on researchers' judgement or what is available in the performance data produced by L2 speakers in specific language tasks. The IC indicators investigated so far are mostly mechanistic markers that index the sequence of interaction. Interaction itself, however, encompasses a much broader range of features that are yet to be theorized in IC assessment. A methodical approach is needed to identify IC tasks that are most relevant to the target L2 group. Based on the identified tasks, IC researchers can further explore what IC indicators are most crucial to success in interaction and theorize these indicators into a more comprehensive IC test construct.

A related research gap in IC assessment research, and in language assessment in general, is that testing specialists have frequently relied on language experts to define the construct of a language test. This can unintentionally weaken the extrapolation inference of a language test to real-world situations as in real life, test-takers' language skills are frequently judged and assessed by non-testing specialists, instead of language teachers or applied linguists (Sato & McNamara, 2019). This issue is particularly pronounced when IC is the trait to be assessed. The ability to interact successfully and appropriately in general everyday-life settings is a skill that everyday-life members in any society can master and need to master. It is not a specialized skill in which only language specialists can claim expertise. To better understand and define IC as a test construct, it can be beneficial to engage with non-testing experts' opinions to understand how IC is understood in everyday-life settings and attained by everyday-life members of a society.

The extrapolative/predictive strength of IC assessment also awaits more investigation. The main function of a language test is to show what is assessed in a testing setting can inform end-users of test-takers' language ability in non-testing settings. IC assessment promises better extrapolation to real-world performance compared to traditional psycholinguistic-individualist speaking tests because IC tests assess test-takers' ability to interact in real-life settings. Such a claim, however, needs to be validated. If an IC test cannot be shown to measure how test-takers behave in everyday real-world situations, it will threaten the inferences we can draw from the test.

Finally, the most contentious point of IC assessment is its relationship with the assessment of linguistic competence (LC, encompassing indicators such as grammar, vocabulary, and pronunciation) as defined in traditional psycholinguistic-individualist speaking frameworks. Is assessing IC the same as assessing LC? If the answer is affirmative, the next logical question is why there is a need for assessing IC since the assessment of LC has long been established and perfected. If the answer is negative, the question that ensues is how IC and LC are different. This is a complex question, and more research is needed to assist us with unpacking the relationship between the two constructs.

In view of the above-mentioned research gaps identified in IC assessment, this book sets out to address these concerns to further our understanding of the practice and value of IC assessment. Chapter 2 offers a literature review of existing IC assessment research and related issues. I provide an exploration of the philosophical underpinnings behind IC assessment, the alternatives to F2F assessment, the relationship between IC and LC, and the practices in test design and rating materials development. An evaluation of existing literature on these issues helps to lay the groundwork for the development of the IC test in this book.

Chapter 3 details the validation framework used in this book to guide the validation of the IC test. In this book, Kane's argument-based framework is used and the assumptions and inferences behind the interpretive argument are detailed in Chapter 3. The project is designed to gather backings to support the assumptions in the framework and the backings will be evaluated in Chapter 7 once all the data are gathered and analysed.

Between Chapter 3 and Chapter 7 is the main focus of this book, which consists of three studies in Chapter 4, Chapter 5, and Chapter 6 that are designed to validate the IC test. Chapter 4 explains the process undertaken in the development of the IC test. Chapter 5 focuses on the design of the rating materials and the specification of the IC construct for the test. Chapter 6 is the main testing study where the test was administered to a large cohort of test-takers to examine how the test and the rating scale function. These three studies generate the evidence needed to validate the test, which is evaluated in the validity argument in Chapter 7.

The last chapter, Chapter 8, summarizes the main findings and contributions from this book, discusses the limitations of this research project, and suggests directions for future research.

Chapter 2 Literature review

Chapter 2 surveys existing literature and research studies that are pertinent to this book. Section 2.1 situates the discussion of interaction in this book in the field of pragmatics, examining the philosophical foundations of three schools of pragmatics theories. An understanding of the philosophical underpinnings behind the theories of interaction is integral to the conceptualization and operationalization of IC in the assessment context. Based on the three schools of pragmatics theories, Section 2.1 proposes a unified framework to situate the assessment of IC, reconciling the tension between the philosophical underpinning for interaction and the real-world constraints of assessment practices.

Section 2.2 establishes computer-mediated communication (CMC) as the platform for the observation of interaction and IC for this book project. The development and assessment of IC in this context are particularly needed and relevant, considering how the global COVID-19 pandemic has shifted much of language learning and testing to the online space. After surveying previous research on CMC and taking into account the unique challenges and opportunities the COVID-19 global pandemic has raised for language testing, Section 2.2 positions the assessment of IC in this book in the CMC environment and raises five considerations in the design of CMC IC tests.

Section 2.3 and Section 2.4 discuss IC as a test construct in detail, with Section 2.3 focusing on theoretical issues pertaining to IC assessment and Section 2.4 on operational issues of IC assessment. In both sections there is a strong emphasis on differentiating IC from LC and on how IC as a construct can be defined to be sufficiently differentiated from LC. A highlight of Section 2.4 is the discussion on the expansion of the IC test construct by incorporating concepts and theories from sociology and philosophy. More specifically, Section 2.4 contends that existing IC assessment research tends to focus on the mechanics of interaction but overlooks the emotional, logical, moral and categorial dimensions of interaction, which should be incorporated into a holistic model of IC.

Sections 2.5 and 2.6 review the literature on designing test tasks and rating materials, both of which are crucial steps in the development of a language test. In Section 2.5 there is a strong focus on how test developers can design IC test tasks that are relevant to test users and that are able to elicit the expected performances as conceived by test developers. This is commonly achieved through the practice of conducting task-based needs analysis (TBNA) to ensure the test tasks are empirically grounded and motivated.

In Section 2.6 the focus is on how rating materials can be developed not from the researchers' perspective but from the language-use domain experts' perspective (Jacoby & McNamara, 1999), and how rating materials can be designed to assist raters with IC rating. Basing the IC construct on domain experts' etic judgement can potentially broaden the IC construct beyond its current mechanistic focus and circumvent the entanglement of IC and LC.

2.1 A philosophical account of interaction

As the 'primordial site of human sociality' (Schegloff, 1992, p. 1296), interaction has fascinated linguists, pragmaticians, sociologists and philosophers alike. This book situates interaction and the ability needed for interaction in the field of pragmatics, though concepts and theories from other academic fields are drawn on to conceptualize interaction. Section 2.1.1 offers a brief summary of pragmatics as a scholarly discipline, highlighting the centrality of interaction to the research agenda of pragmaticians. Section 2.1.2 to Section 2.1.4 examine three schools of pragmatics theories in relation to their takes on the relationship between language and interaction. Naturally occurring conversational data is used to illustrate differing analytic foci of the pragmatics theories surveyed. The connection between pragmatics theories and their philosophical premises is also discussed. Section 2.1.5 provides a unified model combining principles from different pragmatics theories to inform IC assessment in this book. A sound understanding of the theoretical and philosophical underpinnings of interaction is integral to the operationalization of IC as a testing construct.

2.1.1 Interaction and pragmatics

The inception of pragmatics in the late 30s needs to be understood in the broader linguistics research landscape at that time. As the influence of diachronic Neogrammarian historicism waned in the wake of World War I, synchronic Saussurean structuralism was in the ascendency, approaching languages as enclosed systems where phonemes, morphemes, words, clauses, and sentences are hierarchically organized (Bloomfield, 1931; Firth, 1955; Harris, 1951). This structuralist approach has influenced language teaching and assessment ever since and is still traceable in major language tests such as Pearson Language Tests (PTE) where the assessment criteria treat lexis, grammar, and pronunciation as discrete linguistic components (McNamara, 2014; Oller, 1979; Pearson Education, 2020). Though greatly enriching our understanding of language's internal structure, structural linguistics does not extend its analytical attention to the interaction engendered beyond language, providing little information regarding a speaker's

ability to interact in real life. The domineering transformational-generative linguistics after the 50s does not address this issue either, as Chomsky's interest was solely in the competence of an ideal speaker, rather than what functions their real-world language performance fulfils. It is refreshing therefore to see pragmatics, first as a sub-field proposed under semiotics (Morris, 1938), establish its footing between waves of structural linguistics and generative linguistics. Carnap's (1955a, 1955b) advocacy for the study of ordinary language (authentic language performance) and language users, coupled with Wittgenstein's (1953) contextualized view on language use, constituted the philosophical foundation for pragmatics as a distinct field of enquiry. Influenced by works in philosophy of language (Austin, 1962; Grice, 1957; Searle, 1969), pragmaticians positioned themselves as investigators of the connections between speakers, interlocutors, contexts, and interaction.

Having established a shared disciplinary scope, pragmaticians have drawn inspiration from different philosophical traditions and methodological practices to specify interaction, resulting in pragmatics theories of interaction that seem incompatible *prima facie*. The theorization of interaction in this book is influenced by three schools of pragmatics theories, each of which will be inspected, followed by a discussion on how the three schools can be unified to generate an operationalizable account of interaction for the assessment of IC.

2.1.2 An intentionalist perspective on interaction

The first school of pragmatics theories related to interaction is termed intentionalist pragmatics theories in this book, originating in J. L. Austin's conceptualization of speech acts and culminating in his student John Searle's speech act theory. The crux of speech act theory is concisely summarized in the name of Austin's seminal work: *How to do things with words* (Austin, 1962). Speech act theory argues that speaking is not just speaking, it is also doing, and in Austin's theorization there are three layers of doing: a locutionary act, an illocutionary act and a perlocutionary act. A locutionary act can be further dissected into three sub-acts: a phonetic act, a phatic act and a rhetic act. In Austin's definition, a phonetic act describes the action of uttering sounds, a phatic act the production of grammatically meaningful utterances, and a rhetic act the reference the utterance suggests. Below is an excerpt of a naturalistic conversation between two speakers of Mandarin Chinese from my datasets. The conversation took place between a mother and her son. The purpose of Excerpt 1 is to highlight the focus of intentionalist pragmatics theories and their difference from other schools of pragmatics theories.

Excerpt 1 Mother and son conversation

1	Mum	→	ni	zheyang	zuo	hui	jiaoluan	zhengge	jihua	de
			what	this	do	will	disrupt	whole	plan	PRT
			'What you're doing will cause disruption to the overall plan'							
2	Son	→	shenme	jihua						
			What	plan						
			'What plan'							
3	Mum		nide	xuexi	he	gongzuo				
			your	study	and	work				
			'Your study and work'							
4	Son		hao	ba						
			yes	PRT						
			'Ok'							

Note. pragmatics theory data, Chinese, audio-recorded, naturalistic data

Looking at lines 1 and 2, the locutionary act in line 1 is the fact of mum making an utterance, with the phonetic act describing the physical noises this utterance produces, the phatic act reporting the discernible words into which these noises morph, and the rhetic act delineating the literal reference to which these noises point: son's behaviour is disruptive to the overall plan. The illocutionary act is the intended action performed by the utterance. Based on Searle's (1976) classification, multiple plausible readings of line 1 coexist. Mum's intention can be just to state a fact (representative), lodge a complaint against her son's behaviour (expressive), or threaten her son with potential consequences (commissive). The perlocutionary act, which receives relatively less attention in Searle's model, concerns the resultant effect of the speaker's utterance on the hearer, which, in this case, is the son's questioning the meaning of line 1 in line 2. This structuralist approach to action and interaction in speech act theory dissects language into discrete, observable components, bearing a strong resemblance to the componential psycholinguistic approach to speaking and speaking assessment (Levelt, 1989).

The focal point of SA's tripartite analysis is the relationship between locution/speaking and illocution/doing. What drives speaking to doing according to Searle is speaker intention. In fact, Searle has been consistently unequivocal in regard to the intentionalist view he takes. Evidence can be found in the example he gave where a noise becomes a linguistic message when the noise is 'produced

with certain kinds of intention' (Searle, 1969, p. 27), or as he later put it, 'infused with intentionality' (Searle, 1992, p. 7). The intentionalist premise of speech act theory has been a constant source of criticism from Conversation Analysis (CA) researchers and CA-informed discursive pragmaticians (Kasper, 2006; Heritage, 1990). A classic example is Heritage's critique of speech act theory's inability to explain 'oh', an utterance that speakers produce when informed of new information (Heritage, 1990). The uttering of 'oh' is not driven by intention, as Heritage maintained, since no 'vernacular knowledge' (p. 325) of 'oh' existed before researchers inspected this phenomenon using CA. In this book I refrain from taking too critical a stance towards an intentionality assumption behind action since, first, 'oh' is still classifiable under Searle's definition of representatives. Second, though speakers might not have conscious knowledge of 'oh', it is not inconceivable that through socialization they have developed implicit knowledge of issuing 'oh' when wanting to indicate a state of informedness, perhaps without consciousness but nevertheless with intentionality. Psychoanalysis has offered illuminating insight into unconscious intention (Freud, 1956; Siegler, 1967), but a discussion on it is beyond the scope of this book.

In summary, intentionalist pragmatics provides a clear depiction of the relationship between language and action but is less clear about how action forms interaction and what drives the intention behind actions.

2.1.3 A rationalist-utilitarian perspective on interaction

This book defines the second school of pragmatics theories as rationalist-utilitarian pragmatics, typified by Grice's maxims, Sperber and Wilson's relevance theory, and Brown and Levinson's politeness theory. The rationale for drawing a line between speech act theory and the above three rationalist-utilitarian pragmatics theories is that first, speech act theory does not share their utilitarian roots in sociology (Coleman, 1990; Homans, 1961). Second, traditional speech act theory is more intentionalist than rationalist as, for example, in a classic speech act analysis, a person could utter a stream of obscenity with the intention of hurting another person, but later in retrospect they might come to the realization that what they did was not rational. Intentionality can shed light on the mismatch between illocution and locution, but it cannot explain intention itself, as in why intentional behaviour exists in the first place. Rationality, on the other hand, as Føllesdal puts it, 'opens the way to the study of intentionality' (Føllesdal, 1986, p. 115).

Occupying a long-standing position in Western philosophy, rationality is contrasted with passion and appetites in Plato's model of the human soul, revered

as definitory of humankind by Aristotle, and further expounded by modern philosophers such as Descartes, Spinoza, and Leibniz. Rationalism presupposes reasoning power in humans, who are being rational when they exercise this ability to align their behaviours with normative and pro-social expectations. A rationality assumption in the study of human behaviour is tempting. Unable to directly observe human's beliefs, attitudes, and values, researchers can use what philosophers call *a priori* first principles as normative reference points to label behaviours that conform with shared societal expectations as rational, and those that dis-conform as irrational (Davidson, 1980; Hempel, 1961). Though criticism towards such practices abounds in philosophy (Donagan, 1964; Popper, 1957), a rationalist assumption is not uncommon in pragmatics theories, as evidenced by the abovementioned three rationalist-utilitarian pragmatics theories.

The first rationalist-utilitarian pragmatics theory to be inspected is Grice's maxims (1975). Governed by the principle of cooperation, Grice theorized four maxims in relation to quantity, quality, relation, and manner. The rationality assumption implicit in the maxims is that rational speakers tailor their speech to be appropriately informative (quantity), verifiable by evidence (quality), relevant to the context (relation) and perspicuous to the hearer (manner). Though Grice never explicated the reasons or rationale for speaker cooperation, there is reason to believe the Gricean cooperation rests on ethical and moral motivations rather than utilitarian calculations, given the influence of Kantian philosophy on Grice (Grandy, 1989). In other words, the motivation for being cooperative or pro-social for Grice is driven by humans' innate moral compass to be ethical members of society. Relevance theory and politeness theory, on the other hand, are more utilitarian in their philosophical underpinnings. Originating in the works of 18th-century economist Cesare Beccaria, Adam Smith and philosopher Jeremy Bentham, utilitarianism builds on the concept of rationality and holds the view that rational humans, either individually or collectively, act to maximize benefit/pleasure and minimize loss/pain. A utilitarian assumption offers an easy and comfortable explanation for Wilson and Sperber's position on interaction: it is rational for speakers to stay relevant as it increases the clarity of utterance and decreases the effort involved in communication (Wilson & Sperber, 1981; Sperber & Wilson, 1986).

Politeness theory, similarly, has much of its theoretical substratum rooted in rationalism and utilitarianism. Adopting Weber's zweckrational model for the explanation of human behaviour, Brown and Levinson (1987) set out to explicate why *what is said* is not always *what is meant*, a question unaddressed in speech act theory. This endeavour has its origin in the study of social order, most notably in the works of American sociologist Erving Goffman (1967). Face is central to

Goffman's theories of social order and is further analysed into positive face and negative face in politeness theory. Positive face is a person's need to be appreciated by others while negative face is their need to be unimpeded. Threatening another person's face, whether negative or positive, is undesirable since it is in everyone's best interest to preserve one another's face, without which social order is in jeopardy. In other words, driven by utilitarian intents, social order is kept intact by the politeness speakers display to one another. In politeness theory, when implementing an act that might threaten another person's face (a face-threatening act, FTA), a speaker goes through a meticulous rational calculation. Taking line 1 from Excerpt 1 as an example, politeness theory postulates that mum first formulates an intention, decides to instigate an FTA, and chooses to do it on record. In order to mitigate threat to son's positive face, mum adopts redressive action by referring to her son's action instead of him. The three social variables, social distance, power, and rank of imposition posited in politeness theory, here serve as guidance for mum's reasoning. Social distance refers to how close one speaker feels to the other, power the perceived level of authority of one speaker over the other, and rank of imposition the level of difficulty of the action one puts forward to the other speaker. Taking into account the low social distance between herself and son, her higher power over son, and the low ranking of imposition in her criticism/request/statement in Chinese culture, mum brings herself to uttering line 1, which, as made evident by this analysis that adopts politeness theory, is a product of careful means-ends calculation. A detailed explication of the three contextual variables in politeness theory is offered in Section 2.3 where their relevance to IC assessment is discussed.

Though sharing similar philosophical foundations, politeness theory differs from Grice's maxims and relevance theory to the extent that it embeds interaction in its broader socio-cultural context. The speaker and the hearer are positioned relationally by the three social variables and such positionings are always circumscribed by the culture(s) speakers bring to the interaction. In this sense, politeness theory signals the social turn that pragmatics as a discipline has taken from its more structuralist and psycholinguistic starting point in philosophy of language.

2.1.4 An empiricist-interactional perspective on interaction

The third school of pragmatics theories is defined by its empiricist-interactional approach to interaction, typified by the practice of Conversation Analysis (CA) (Sacks et al., 1974). Different from the previous two schools that draw strongly from concepts in philosophy of language, CA follows the sociological

tradition and situates its investigation in social organization (Garfinkel, 1967; Goffman, 1955; Habermas, 1981; Parsons, 1937). Though CA shares its founding principles in the Goffmanian interactional order as does politeness theory, CA adopts an epistemological stance and analytic policy that are decidedly empiricist (Buttny, 1993; Schegloff, 1996). Compared with the previous two schools of pragmatics theories, CA remains agnostic on the intent behind action (Heritage, 1990), rejects *a priori* assumptions on practical/utilitarian reasoning, and refuses presupposed macro sociocultural context in its unmotivated looking at conversational data. A reanalysis of the same mother-son data in CA fashion in Excerpt 2 can illuminate the differences.

Excerpt 2 Mother and son conversation CA-transcribed

```
1  Mum  →  ni  ↑zheyang zuo hui  jiaoluan zhengge jihua [de=
           what this    do  will disrupt  whole   plan   PRT
           'What you're doing will disrupt the overall plan'
2  Son  →                                               [=shenme jihua
                                                          What   plan
           'What plan'
3  Mum     nide   xuexi   he      gongzuo
           your   study   and     work
           'Your study and work'
           (2.2)
4  Son     >°hao:     ba°
           yes       PRT
           'Ok'
```

Note. pragmatics theory data, Chinese, audio-recorded, naturalistic data

Following a CA analysis, in line 1 mum offers a negative assessment of son's behaviour, but before she finishes, son's repair attempt is launched and overlaps with her line 1. Son's closed other-initiated repair questions the definition of 'plan', leading to mum's self-repair in line 3. A noticeable lapse of 2.2 seconds follows before son finally issues an agreement token in line 4.

Based on Excerpt 2, the two most noteworthy features of CA-informed empiricist-interactional pragmatics are detailed here. The first is CA's empiricist epistemology. As influential as rationalism in Western philosophy, empiricism

finds its origins in Aristotelian philosophy and the works by classical British empiricists such as John Locke and David Hume. Vigorously maintaining that knowledge derives from observable experience and evidence (Schegloff, 1996), CA analysts examine the visible sequential structure of conversation data without mentalizing the psychological states of speakers. In the case of line 1, it is common in, for example, speech act theory, to prescribe or speculate mum's true intent behind her locutionary act, be it warning, suggesting, or reprimanding. Such practice, however, is strictly forbidden in CA. Even the temporal information gleaned from CA transcription is treated with caution in CA analysis in terms of its generalizability. The unusually long lapse between lines 3 and 4 suggests itself as a candidate phenomenon, which means a phenomenon worthy of generalization in CA parlance (ten Have, 2007). However, until verified by a larger collection of similar specimens, this phenomenon remains localized to this very specific incident and cannot be claimed to be applicable to other situations (Heritage, 2011; ten Have, 2007; Wu, 2016). Empiricist-interactional pragmaticians are suspicious of any premature extrapolation of one single specimen to generalized statements such as 'son is using silence to show his displeasure'. Arguments such as 'son has to acquiesce in line 4 because children need to show deference to their parents in Chinese culture' are equally questionable in CA as characterization of sociocultural factors or contextual factors such as power should not be assumed relevant unless proven empirically in data. This contrasts sharply with the practice commonly seen in politeness theory.

The second most noteworthy feature of the empiricist-interactional approach to interaction is its interactional perspective on conversation. Of all the pragmatics theories discussed so far, CA-informed pragmatics (discursive pragmatics as termed in Kasper, 2006) is richest in its encapsulation of the contingent, interactional nature of conversation. As we can see in the vignette in Excerpt 2, if mum's intention in line 1 is to seek agreement, this agreement is not immediately available in line 2. In fact, only after a repair, a clarification and a lapse is it finally offered in line 4. Austin's perlocutionary act, Grice's relation maxim and Sperber and Wilson's concept of relevance pale in comparison in their ability to capture such fine interactional details. Searle's speech act theory, in contrast, is at the other end of the spectrum, emphatically unit-act-focused and disinterested in applying itself to interaction. As Searle insisted in his 1992 article, once his intentionality-infused speech act is delivered, silence ensues and the speaker and the hearer 'go their separate ways' (p. 14). Politeness theory, though initially formulated on the single act model, is more amenable to sequentialization (see Introduction to the reissue in Brown & Levinson, 1987).

However, sequence is not traditionally a central concern in research that adopts the politeness theory framework.

2.1.5 A unified account of interaction for assessment

After the previous sections have situated interaction in pragmatics and reviewed three approaches to interaction in pragmatics theories, this section aims to explain why none of the above-reviewed approaches alone is enough to conceptualize interaction in the assessment context and why a unified approach is necessary. Let us first inspect again how the three approaches map onto the interconnected relationships among language, action, and interaction. If we zoom in on the connection between language and action, speech act theory offers a sound explication of *what is said* (language) is not usually *what is meant* (action), and *what is meant* (action) is driven by intention. Though etically there can be multiple readings of the intention behind an utterance (action), it is difficult to falsify the existence of speaker intention (see the analysis of Ochs, 1984 in Brown & Levinson, 1987). The postulation of speaker intention is necessary for the assessment context as speakers' speech is perceived as ratable evidence indicative of latent competence. Assessment will be impossible if speakers' performance can be perceived as unintentional or free of speakers' will. However, speech act theory alone is unable to capture the complexity of interaction as its main focus is on the relationship between speech and action at a localized single-act level.

If we zoom out to look at how actions transform into interaction between speaker and hearer, we notice interactional features such as turn-taking and recipient design as in the analysis in Section 2.1.4, which can be empirically observed and validated through CA. However, the empiricist tradition in CA encourages a focus on naturalistic interaction that takes place *in the wild* (Hellermann et al., 2019), a term first proposed in Hutchins (1995) and frequently used in CA research to describe interactional activities that take place outside controlled settings. The *in-the-wild* preference in the empiricist CA tradition is a condition that large-scale standardized assessment is unable to meet. If assessment specialists are to standardize interaction in a testable manner, they have to impose broader contextual parameters such as power, distance, and imposition in politeness theory to regulate the context of the interaction and the relationship between interactants. This abstracts interaction *in the wild* to one that is observable and regulatable in a controlled assessment context.

Another limitation of adopting an orthodox empiricist-interactional approach in the assessment context is that the empiricist-interactional approach is disinterested in making value judgement regarding interactants and interaction,

which is misaligned with the principle of assessment. The ability to interact appropriately and successfully is a social achievement via mutual endeavours by the speaker and the hearer so as to preserve social solidarity and social order. The orthodox CA-based empiricist-interactional approach tends to only focus on the emic description of such abilities and achievements, instead of assessing them from an etic, testing perspective.

A third limitation in a solely empiricist-interactional approach is that CA itself is unable to explain the cooperative and pro-social nature of interaction that it often notices in interactional data. The rationalist-utilitarian approach, on the other hand, offers an explanation for the pro-sociality of human beings. Levinson attributed it to inherent concerns of face and FTAs (Levinson, 1983), explaining CA findings away through the lens of politeness theory, while Grice postulated it on the inherently cooperative nature in humans, possibly due to influences from Kantian moral motivations. Clayman (2002), when explicating the CA concepts of *preferred* and *dispreferred responses*, resorted to politeness theory and the Goffmanian concept of face to justify humans' needs for solidarity and cooperation. This further underscores that rationality and utilitarianism can be used as reference points for the explication of human interaction, calling for unity between rationalism and empiricism in their respective pursuit of knowledge of human interaction. This is not a maverick, reckless proposition since Kantian philosophy has long argued for a unified look at rationalism and empiricism. Methodological differences can be reconciled to generate a more robust framework for the specific research context and questions on hand. Drawing on CA, interactional sociolinguistics, politeness theory, critical discourse analysis, and discursive psychology, Stubbe et al. (2003) illustrated how methodological differences in the five distinct approaches can be reconciled to generate complementary findings on interaction from the same nine-minute audio recording. The authors argued:

> Each approach therefore provides a slightly different lens with which to examine the same interaction, highlighting different kinds of context and respective features. These are not necessarily in conflict with one another (though in some cases the analyses and/ or the theoretical assumptions underlying them are difficult to reconcile); rather, they are complementary in many ways, with each approach capable of generating its own useful insights into what is going on in the interaction, with the proviso that the framework adopted needs to be a good match for the research questions being asked. (p. 380)

In view of this, this book adopts a unified approach combining intentionalist, rationalist-utilitarian and empiricist-interactional pragmatics to operationalize the assessment of interaction at a philosophical level. Below are nine principles that

explicate the philosophical premises for the theorization and operationalization of interaction in this book.

1. The unified philosophical approach to interaction in this book posits that the ability to interact is, on one hand, the ability to intentionally launch actions via speech (intentionalist), and on the other hand, the ability to make action recognizable to other interactants and attain interactional goals (rationalist-utilitarian). The observation of IC needs to be verified in interaction, through the inspection of interactional data (empiricist-interactional).
2. IC is conceptualized as an individual ability in test-takers. Though interaction is co-constructed, IC is an ability that test-takers can claim and demonstrate through interaction, making the assessment of one's individual IC and awarding an individual IC score possible.
3. Test-taker speech is viewed as indexical of their IC because test-takers are considered to have the intention to be in control of their speech and to demonstrate their IC (the intentionalist assumption).
4. Test-takers are postulated to be rational insofar as their speech reveals their values, beliefs, attitudes, and reasoning surrounding interaction. Test-takers are pre-supposed with the intention to be cooperative and pro-social because they possess the reasoning power to appreciate that it is rational to maintain strong interpersonal relations through interaction. This makes it possible to make value judgement regarding test-takers' interactional conduct and their handling of the interaction.
5. Though intentionalist and rationalist-utilitarian assumptions are attributed to test-takers, in this book I do not base the analysis of interaction on atomized language (intentionalist) or means-end calculation (rationalist-utilitarian). Instead, I adopt a data-driven empirical approach in the analysis of test-taker performance. In other words, following the tradition in empiricist-interactional pragmatics, test-taker performance is analysed in its interactional environment without speculation of their mental states. The normative intentionalist and rationalist-utilitarian assumptions about test-takers still exist, nonetheless.
6. Contextual variables of interaction such as the three variables in politeness theories (power, social distance, and rank of imposition) are admitted in the operationalization of the test so as to standardize the interactional scenarios in the test.
7. Congruent with the tradition in CA, broader socio-cultural factors surrounding interaction are treated as unverified unless they are made relevant by interactants. In other words, social or cultural explanations are

not evoked to explain interactants' behaviour *a priori*, for example, 'the interactant is speaking in this manner because it is common to do so in the Chinese culture'. The discussion of socio-cultural factors needs to be validated through interactant discourse.

8. The task scenarios in the test can be perceived as leading the test-takers to adopt certain speech acts (or social actions in CA parlance) for the purpose of standardization in testing, but such a relationship is not causal. The focus of assessment is the quality of interaction, instead of what speech acts/social actions are launched.
9. Controlled speech performance, such as role-play, is permissible in assessment settings for practicality concerns. Role-play data is considered valid, empirical evidence indicative of test-taker IC. This approach to data collection differs from orthodox empiricist-interactional CA, where a strong focus is put on the elicitation of naturalistic data (ten Have, 2007). It is nevertheless considered a necessary logistic compromise for language assessment.

The nine principles explain the philosophical underpinnings of interaction and the assessment of interaction in this book. This operational definition of interaction does not reject or endorse any particular pragmatics theory or align solely with any of the three pragmatic traditions discussed. Instead, following the tradition in Kantian philosophy, which unifies rationalism and empiricism, this book draws on the strengths of different theories and defines a unified philosophical basis for the operationalization of the assessment of interaction that is consistent with the philosophy and practice of language assessment.

2.2 Interaction in computer-mediated communication

Having investigated the philosophical bases for conceptualizing interaction in the assessment context in Section 2.1, Section 2.2 focuses on the assessment context where interaction takes place, and in this book project, it is computer-mediated communication (CMC). The rationale for adopting CMC as a site for observing L2 speakers' interaction is detailed in the following sections. Existing research on L2 pragmatics and interaction in CMC is also reviewed.

2.2.1 CMC and L2-speaker interaction

Though commercial Internet services emerged in the late 1980s and only became widely accessible at the start of this century, the use of CMC has proliferated and permeated through every form of interpersonal interaction. This shift in communication patterns has naturally drawn the attention of language educators

and assessors given that communication is the primary platform for language acquisition (Crystal, 2011; González-Lloret, 2011). The COVID-19 global pandemic further spotlighted the efficacy, convenience, and necessity of CMC in language teaching (Bozkurt et al., 2020; Junn, 2021; Kohnke & Moorhouse, 2020; Moorhouse & Beaumont, 2020; Moorhouse et al., 2021) and language assessment (Isbell & Kremmel, 2020; Ockey, 2021). In view of this, language educators and assessors have started to systematically interrogate the practicality and benefit of language development and assessment in the CMC realm (Gruba et al., 2020; Oh, 2020).

Positioning the study of interaction in CMC is, however, not an easy endeavour. Compared to traditional face-to-face communication (F2FC), the various modes and sites of CMC pose unique challenges to the observation of L2 interaction on CMC platforms. Not only are there textual, audio and video modes of CMC, but there are also multiple sites where CMC takes place: smartphone applications, social networking websites, online video games, and online discussion forums. If CMC as a medium of interaction is to be operationalized in the assessment context, assessment specialists need to specify the modes and sites of CMC to be adopted. This decision should not only be informed by conventional wisdom of what modes and sites of CMC are compatible with language testing, but also by research that specifically investigates the differences in language use among different CMC modes and sites. So far, we are not seeing enough studies comparing different modes of CMC in second language contexts as most research focuses on juxtaposing one mode of CMC with F2FC (Kim, 2017; Loewen & Isbell, 2017; Rouhshad et al., 2016; Ziegler, 2016). Hung and Higgins (2016) is a rare exception that compared learning gains between textual CMC (MSN Messenger) with video CMC (Skype), concluding that textual CMC provided more opportunities for learning target forms while video CMC was more effective in improving fluency and pronunciation.

In terms of the CMC modes investigated in the field of L2 pragmatics, it is regrettable that the modes examined so far are quite limited. Computer-based email writing in academic and business settings is by far the most widely researched CMC mode in L2 pragmatics (Lan & MacGregor, 2010; Nguyen, 2018). While general SLA has already started to explore audio and video modes of CMC on Skype and other social networking platforms (Dooly & Sadler, 2016; Dugartsyrenova & Sardegna, 2017; Terhune, 2016), we are yet to see this trend mirrored in L2 pragmatics.

Assessment research in CMC L2 pragmatics has a similar penchant for email, whether in research-based studies (Chen & Liu, 2016; Youn, 2014) or commercial testing settings (CambridgeESOL, 2012; Laughlin et al., 2015). The

ability to write appropriate emails undoubtedly is a useful skill for L2 speakers since pragmalinguistic and sociopragmatic blunders in emails can attract severe sanctions (Taguchi, 2012) and emails remain an indispensable medium for formal business communication (CambridgeESOL, 2012). However, a broadening of the research scope in CMC-L2 pragmatics is clearly in order. Prompted by the COVID-19 pandemic, there is strong motivation to assess L2 speakers' ability to interact successfully and appropriately in spoken modes of CMC.

2.2.2 An empiricist-interactional approach to CMC

The previous section explains why CMC offers a fertile ground for language educators and assessors to develop and assess students' ability to interact in the target language, especially after a global pandemic. However, before computer-mediated interaction can be observed or assessed, researchers need to first understand how computer-mediated interaction unfolds. Aligning with the empiricist-interactional approach to interaction taken in this book, here I argue that CA methodology can offer valuable insight into the unfolding of computer-mediated interaction. The advantages of employing CA in the study of CMC are clear: CA's rigorous analytic focus on minute details can cast light on the systematic procedures hidden behind seemingly random and chaotic conduct in CMC. Negretti (1999) marked the first attempt in this area by comparing the turn-taking, opening, and closing of online chat between L1 English speakers and L2 English speakers, noticing that both L1 and L2 speakers were quick to reconfigure their F2FC methods to the constraints of online communication. Similar attempts after Negretti (1999) are sporadic, which leads González-Lloret (2015) to lament the lack of visibility of CA for CMC in the authoritative *Wiley Encyclopedia of Applied Linguistics* (Chapelle, 2013) and the *Wiley Blackwell Handbook of Conversation Analysis* (Sidnell & Stivers, 2013). However, in recent years the tide is turning (Tudini & Liddicoat, 2017) as we are seeing a growing number of studies dedicated to CA-for-CMC in video chats (Jenks, 2014; Licoppe & Morel, 2012), texting (Abe & Roever, 2019, 2020; Hutchby & Tanna, 2008; Tudini, 2010), and online group tasks (Sert & Balaman, 2018). Journals have also dedicated special issues to the exploration of CA in the study of L2 interaction and interactional competence (IC) in computer-mediated environments (see Dai et al., 2022a for an edited special issue on L2 IC in the online space).

Given CA's primary interest in uncovering the sequential organization of social actions, a central question for CA-for-CMC researchers is how much of the insight traditional CA research has gleaned from synchronous communication (e.g., F2FC and telephone calls) is transferable to asynchronous CMC (e.g.,

texting and voice-messaging). One of the earliest attempts in this regard comes from Schönfeldt and Golato (2003). The researchers selected repair as a candidate phenomenon for the comparison between F2FC and online text chat, concluding that interlocutors drew methods from basic repair mechanisms (Schegloff et al., 1977) and adapted them to the specific parameters and constraints of online web chats. Certain repair procedures such as using silence to prompt the previous speaker to initiate self-repair cannot be realized in text chat, as silence between messages/turns is not interpreted in the same way as in F2FC. A five-minute gap between messages can be because the other interlocutor is engaged elsewhere rather than suggestive of a dispreferred second part, though unexpected gaps in otherwise quasi-synchronous text-messaging sequences can certainly be oriented to as such. In summary, Schönfeldt and Golato (2003) suggest that text chat shares many similarities with F2FC, though noticeable differences in interactional methods also exist.

Some non-CA CMC researchers have gone one step further than Schönfeldt and Golato (2003) by voicing against what they consider a contemporary exceptionalism fallacy in our understanding of CMC. Herring (2013) and Thurlow and Mroczek (2011) argued that CMC was not as different as we would like to think in terms of interactional methods. The patterns of interaction in CMC are still embedded in ordinary F2FC and interlocutors constantly orient to their most familiar F2FC methods as normative reference points. Along this line of argument Locher (2014) maintained that when we revisit the speaking mnemonic from Hymes (1974), we realize it is the same human beings that are using computer-mediated media to communicate. Therefore, human interlocutors are restricted by the same participation structures and regulated by the same conversational norms as in F2FC. González-Lloret (2015), on the contrary, holds an opposite view that different media of CMC need to be examined in their own right just like how telephone conversations were first investigated in CA as an independent system (Schegloff, 1979). She contends that the fact that certain methods or features of CMC are different from the ones of F2FC indicates that CMC does not always conform to the norms in ordinary conversation and that CMC is highly medium-dependent (e.g., text-messaging has a system of its own compared to video chats). Such conflicts in opinions are largely due to the paucity of bottom-up empiricist research comparing interactional methods across different CMC modes and between CMC and F2FC. A solution to this query can be to investigate if the interactional patterns and methods gleaned from F2FC can be applied to interaction in different CMC modes. In this book, I adopt an empiricist bottom-up approach by starting with

CMC interactional data and examining if insight from F2FC CA can be applied to the CMC environment.

2.2.3 Five CMC considerations for test design

Having positioned the study of interaction in the CMC context for this book and surveyed existing research relevant to this topic, here I will detail five considerations that inform the operationalization of assessing interaction via CMC in this book project.

1. **Expansion** of modes for the investigation of L2 IC. The study of L2 interaction needs to go beyond F2FC and venture into CMC. Existing research on L2 speakers' IC has mainly engaged with F2FC, either in classroom-based naturalistic studies (Cekaite, 2007; Hellermann, 2008) or research-based role-plays (Youn, 2013; Al-Gahtani & Roever, 2014, 2018). Though we are seeing a growing awareness of situating L2 IC in CMC, such initiatives need to be taken up and expanded, especially considering how the global pandemic has made CMC not a choice but a must in many language education and assessment contexts.
2. **Careful selection** of CMC sites for investigation. The assessment platforms of CMC need to match with the constraints of assessment and L2 speaker groups' needs and practices. Though every CMC site can generate insight into CMC interaction, testing researchers should prioritize sites that can provide the most authentic interaction that is able to be captured in large-scale language testing. Following this argument, though online video games are a promising site to inspect the development of interactional skills in CMC contexts (see research on Final Fantasy in Piirainen-Marsh & Tinio, 2014 and Second Life in Pojanapunya & Jaroenkitboworn, 2011), these CMC sites are not most easily adaptable to large-scale standardized language tests. Other CMC sites that have the potential for testing purposes and have been investigated so far include MSN (Hung & Higgins, 2016), Skype chatroom (Jenks & Brandt, 2013), radio station chatroom (Schönfeldt & Golato, 2003), Livemocha (Gonzales, 2013) and Google Hangouts (Sert & Balaman, 2018). Which of these sites to choose for the medium of assessment needs to be based on the particular needs and common practices in the test-taker groups so that the assessment tools are authentic and relevant to L2 speakers' learning and communication practices.
3. Prioritization of **non-textual CMC modes.** More studies need to be conducted on non-textual CMC modes in L2 pragmatics. Though the continuing interest

and research output in text-based modes such as emails (Nguyen, 2018) and text chats (Abe, 2018) are certainly fruitful, research in non-textual CMC should be encouraged. Surveying the change in communication methods among 1,374 American adults during the pandemic, Nguyen et al. (2020) reported that 43 % of the respondents were sending more text messages, which is a textual CMC mode. Meanwhile, 36 % of respondents reported more usage of voice calls and 30 % of respondents used more video calls, both of which are non-textual CMC modes. It is clear that the pandemic promoted the use of not only textual CMC but also non-textual CMC. Though Nguyen et al. (2020) did not differentiate L1 and L2 speakers in their 1,374 participants, the change in CMC frequency noted in their study is likely to hold true for L2 speakers as well, pointing to the need for more attention in L2 pragmatics research on non-textual CMC modes. Furthermore, mode-dependent interactional features and methods have been noticed in non-textual CMC research on L1 speakers (e.g., video chats in Licoppe & Morel, 2012 and mobile phone calls in Arminen & Weilenmann, 2009). IC researchers can similarly investigate if such interactional patterns exist for L2 speakers and how these patterns can be assessed in testing settings.

4. **Collation and comparison** of performances in different non-textual CMC modes. Building on the previous consideration, L2 speakers' interaction via different non-textual CMC modes needs to be collected to offer comprehensive coverage of the test construct. Let us take the popular mobile-phone-based interaction application WeChat as an example. WeChat is frequently used among L1 and L2 Chinese communities, allowing users to send voice messages, respond to voice messages and conduct live video chats. These different interactional modes represent potentially different aspects of the IC test construct and measure the construct at different levels. If an IC test only measures the ability to send voice messages, which is asynchronistic and represents a narrower construct of IC, the test results cannot reliably inform stakeholders of test-takers' ability to conduct online real-time interaction such as video chats. In view of this, test designers need to select different CMC modes so as to make the test results more generalizable and informative for end-users.
5. The foregrounding of **L2 speakers** in the investigation of interaction. The last consideration highlights the necessity of foregrounding L2 interaction in the burgeoning CA-for-CMC literature. Despite the spotlight on CA-for-CMC (for a special issue see Antaki, 2016), L2 speakers form a largely under-researched cohort of CMC users. With CMC researchers preoccupied with L1 speakers and CA-for-SLA researchers privileging traditional F2FC, how

> L2 speakers experience and negotiate the ever-changing world of CMC is largely unknown, which requires immediate attention given the pedagogical and assessment implications COVID-19 has highlighted for us.

Based on previous research on interaction in CMC, the abovementioned five considerations have guided and shaped the design of the IC assessment tool in CMC for this book. Details on how these considerations translate to specifications in test design can be found in Section 4.2 in Chapter 4.

2.3 Defining an IC construct: A theoretical discussion

Having explicated the philosophical foundations of studying interaction and situated second language interaction in the CMC context for this book in the two previous sections, this section narrows the discussion on IC to the language testing realm and explains the theoretical challenges in IC assessment.

2.3.1 A brief history of IC

The discussion of interaction in Sections 2.1 and 2.2 approaches interaction in its general sense, which is synonymous with interpersonal communication. The term interactional competence (IC) has been used loosely to describe the ability in interaction or interpersonal communication, although IC is yet to be formally defined. From Section 2.3 onwards, the concept of the ability to interact will be specified in language testing terminology and defined formally as IC.

The origin of the concept of IC in the second language context is often credited to Kramsch (1986). In response to the communicative language teaching movement in Europe, Kramsch emphasized IC, which embodied the functional purposes foreign language teaching in the US should aim for. Further furnished by communicative competence theories (Hymes, 1967, 1972; Canale & Swain, 1980; Celce-Murcia et al., 1995; Celce-Murcia, 2007) and analytic tools afforded by CA, IC has been reconceptualized as the competence that speakers demonstrate when they orient to one another's previous utterances, produce sequentially appropriate responses and co-construct social actions (Hall & Pekarek Dohler, 2011; Pekarek Doehler & Berger, 2016; Roever & Dai, 2021; Roever & Ikeda, 2021). Although IC can also be defined as a co-constructed competence residing in the interaction and distributed between speakers (Mehan, 1979; Young, 2011), following the principles of theorizing interaction for this book as laid out in Section 2.1.5, IC in this book project is defined as a competence that an individual/test-taker can claim and display in an assessment context.

Before CA's empiricist-interactional tradition started exerting its influence on the theorization of IC, the assessment of L2 speakers' ability to communicate, or more precisely, their pragmatic competence, had largely adopted the intentionalist and rationalist-utilitarian traditions in pragmatics theories (Grice, 1957; Sperber & Wilson, 1986; Brown & Levinson, 1987). Focusing on L2 Chinese, Taguchi (2015) examined 14 empirical L2 pragmatics studies; findings showed that all the surveyed studies looked at single-act constructions such as speech acts, implicatures and routinised expressions, using non-interactive research instruments such as discourse completion tasks, multiple choice questions and judgment tasks. This testifies to the enduring influence of traditional intentionalist and rationalist-utilitarian pragmatics frameworks.

Though these single-act cognitive pragmatics theories can offer insight into the meaning-action relationship in their own rights, as discussed in Section 2.1, relying on intentionalist and rationalist-utilitarian theories alone can under-theorize both interaction and interactional contexts as these theories assume a one-to-one relationship between utterance and action, without sufficient recognition of the contingent nature of interaction. Such a propensity directly maps onto assessment tools that are commonly adopted in L2 pragmatics research. Instruments such as discourse completion tasks and sentence judgment tasks take an exogenous approach to interaction, which, as Golato (2003) demonstrated, results in measuring speakers' metapragmatic knowledge instead of their online interactional behaviours.

Dissatisfied with previous theorizations of interaction and assessment tools of interaction, pragmaticians have turned to sociological theories, most notably, CA, for a better understanding of interaction. This led to the theorization of the concept of IC, drawing on the empiricist-interactional approach to interaction discussed in Section 2.1.4. The last ten years have seen great improvement in understanding of L2 IC from a developmental perspective (for overviews see Kasper & Wagner, 2011; Lee & Hellermann, 2014; Pekarek Doehler & Pochon-Berger, 2015; Pekarek Doehler et al., 2018). Longitudinal developmental IC studies have looked at how L2 speakers' IC develops over time in terms of turn organization in disagreement (Bardovi-Harlig & Salsbury, 2004), multiparty classroom talk (Cekaite, 2007), task opening, storytelling and disengagement in classroom activities (Hellermann, 2007; 2008) and story opening in host family settings (Pekarek Doehler & Berger, 2016).

In contrast, cross-sectional developmental IC studies examine how L2 speakers at differing L2 proficiency levels interact differently, using IC measures informed by CA findings. Conversational features investigated so far include the range of methods learners mobilize to implement disagreement (Pekarek

Doehler & Pochon-Berger, 2011), the presence or absence of preliminary moves in request (Al-Gahtani & Roever, 2012), clear or unclear dispreference markings (Al-Gahtani & Roever, 2014), features in the construction of agreement and disagreement (Zhang, 2016) and recipiency design (Al-Gahtani & Roever, 2018).

As the concept of IC flourishes in second language teaching and education in general (see Dai et al., 2022b for a recent special issue on this topic), there has also been a strong interest in the assessment of IC (Salaberry & Burch, 2021). This is evidenced by a special issue of the journal *Language Testing* (Plough et al., 2018) devoted entirely to IC, a special issue on the employment of CA in assessing IC of the journal *Papers in Language Testing Assessment* (Youn & Burch, 2020) and the prominent role of IC assessment in the special issue of the journal *Language Assessment Quarterly* on speaking assessment (Lim, 2018). Despite the growing interest in IC assessment, to this date no major language test assesses IC. Assessing IC in standardized testing poses different challenges to IC researchers compared to studying IC from a developmental perspective. Testing is an expensive practice that has logistic, practical, and economic implications. Given that major language tests are already assessing speaking and have shown to be reliably assessing it, it is understandable for testing specialists and testing companies to wonder what the value-add is in assessing IC, if assessing IC is the same as assessing linguistic competence (LC, including componential linguistic control over lexicon, grammar, pronunciation and so on), which is commonly assessed in speaking tests. In other words, the questions IC assessment researchers face are whether IC needs to be assessed and what additional benefits IC assessment brings, compared to the established practice of assessing LC in speaking tests. The rest of Section 2.3 focuses on discussing why the absence of IC assessment is problematic, why simply assessing LC does not give information about test-takers' IC, and how an IC construct differs from an LC construct.

2.3.2 Assessing IC

In the widely used argument-based approach to assessment (Kane, 2006, 2012), the purpose of a test is to generate desirable consequences, or more specifically, to provide information in the form of scores about the strength of an attribute of interest in test-takers, such as language skills, to enable decisions about these test-takers (see Chapelle et al., 2008; 2010, for examples). In the case of language tests, decisions informed by language test scores might include such high-stakes decisions as admission to a foreign university, suitability for practising as a medical professional, or permission to settle permanently in the host country. A test is arguably doing its job well if it enables good decisions, e.g., foreign

medical graduates really do have the necessary language skills to communicate with patients and fellow professionals.

In the overwhelmingly typical case where test-takers' real-world language use involves interacting with others, it is reasonable that their ability to do so should be a core part of the information gathered on their language ability. Language is a tool to make private thoughts public and visible (audible) to others, obtain access to their private thoughts, and thereby enable coordinated social actions. As discussed in Section 2.3.1, language users' ability to deploy language for accomplishing social actions has been conceptualized as their IC, defined by Hall and Pekarek Doehler (2011) as the 'ability to accomplish meaningful social actions, to respond to co-participants' previous actions and to make recognizable for others what our actions are and how these relate to their own actions' (p. 2). This includes behaviours by which social roles are enacted in a given context, context-specific ways of organizing turn-taking and communicative practices, as well as the use of linguistic and non-linguistic resources to accomplish these goals. While writing and reading are also to some degree forms of interaction, as the review on L2 pragmatics and interaction in Section 2.2 indicates, the immediate coordination of social actions between people on a moment-by-moment basis relies most strongly on speaking and listening. Language users' social actions are specifically designed with regard to their recipient (Drew, 2013) and every utterance accomplishes a social action that provides a context for the interlocutor's subsequent actions (Heritage, 1984): this is the core meaning of inter-action.

It seems logical that language tests should mirror this interactive use of language in order to obtain a representative picture of what the user can do in the real world and allow extrapolation from the sample of language use situations in the test to real-world language use. Alas, they do not. Language tests frequently do not include interaction in a second language, and where they do, they do not assess it, instead just using the resulting talk as a language sample to be rated on non-interactive criteria. For example, TOEFL and PTE contain monologic speaking tasks where test-takers react to prompts and input materials without engaging with an interlocutor. Tests like the ACTFL OPI and the IELTS speaking test involve a live interlocutor but their rating scales do not include measures of interactional abilities, and their main purpose is simply the elicitation of samples of spoken language. Various fine-grade empirical studies have demonstrated the non-equivalence of interview-based speaking tests and natural conversation (e.g., Lazaraton, 1992; Johnson, 2001; van Lier, 1989).

The Cambridge English scales for assessing speaking performance (Cambridge English, n.d.) used for the Cambridge main suite exams go a step

further and contain a rubric for interactive communication. However, this rubric is strongly influenced by the test format and focuses on maintaining interaction, responding to the interlocutor, and linking contributions to the interlocutor's, thereby likely abbreviating the construct of interaction. Finally, the Common European Framework of Reference (CEFR; Council of Europe, 2001, 2018) has a number of scales for different interactive activities (casual conversation, interview, negotiation) and 'interaction strategies' (taking the floor, ensuring understanding, repair) but pays little attention to social action implementation (Roever & Dai, 2021) and social role enactment (Dai & Davey, 2022, 2023). Tests of languages other than English, such as the Test of German as a Foreign Language (TestDaF) or the Chinese Standard Exam (Hanyu Shuiping Kaoshi, HSK) also do not score IC.

This lack of attention to IC means that most language tests privilege speaking over talking, to use Roever and Kasper's (2018) parlance. Roever and Kasper (2018) viewed talking as interactive language use, including designing utterances for a specific interlocutor and comprehending implied social actions. By contrast, speaking is simply a monologic response to a stimulus not designed to achieve a social action vis-à-vis an interlocutor. While people usually talk to others and rarely just speak to no one, the latter is exactly what the vast majority of language tests assesses. Revisiting the philosophical discussion on interaction in Section 2.1, it is clear that traditionally the assessment of speaking has a structuralist, atomized view towards interaction, focusing solely on the linguistic components of speaker utterance, with little consideration of the inter-action that utterances engender in the natural, interactive environment of interpersonal communication. This highlights the difference between assessing IC/talking and assessing LC/speaking.

Given that large-scale, high-stakes language tests do not assess what people use language for, there is a serious risk that their results are flawed and that construct underrepresentation threatens the defensibility of decisions and conclusions based on scores (Messick, 1989). There would be no problem if speaking is the same as talking, i.e., if people who are good at speaking/LC are invariably good at talking/IC, and people who are good at talking/IC are invariably good at speaking/LC. In the following deliberations I will first show that speaking/LC is in fact not the same as talking/IC, and then I will discuss ways of making language tests more representative of what people do with language when talking.

2.3.3 Differentiating speaking/LC and talking/IC

Speaking ability following the classic model by Levelt (1989, 1999), which explicitly informs tests such as PTE, requires fast access to vocabulary,

automatization of grammatical knowledge, and high-speed phonetic encoding and articulation. This corresponds to the traditional definition of LC, which is treated as dissectible into a host of separate components such as vocabulary, grammar, and prosodic features. Such components are assessed independently of each other and isolated from the contexts where they occur (see Section 2.1.1 for a review of the history of structuralism in linguistics). In other words, the assessment of speaking/LC represents a structuralist view in measuring speech production, without reference to what action the speech achieves in its interactional environment or what inter-action the speech effects between speaker and hearer. It is fundamentally an assessment of speech without relating it to its real-world implications.

This structuralist, decontextualized approach to assessing speaking is not uncommon in existing testing practices. The privileging of fluent, smooth, easily comprehensible speech is captured well in the test construct of L2 facility, which Bernstein et al. (2010) freely admitted 'provides a measure of performance with the language without reference to any specific domain of use' (p. 356). Similarly, the IELTS band descriptors (IELTS, n.d.) for the speaking test rate test-taker performance in four categories: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation, all of which are markers of LC. There is no mention of interactional abilities in the IELTS speaking rubric, even though IELTS speaking involves a face-to-face interview, and research exists on the interactional construction of this interview (Brown, 2003; Seedhouse & Egbert, 2006; Seedhouse & Nakatsuhara, 2018).

Roever and Kasper (2018) termed this the psycholinguistic view of speaking, and it sees speaking ability as consisting of a set of linguistic components that can be assessed without reference to their conversational use. Still, it assumes that such an assessment will provide useful information since these abilities are required across a wide variety of contexts. This is probably not entirely unreasonable: if speakers have ready access to vocabulary and can implement a broad range of grammatical functions under real-time conditions, this will help them under any circumstances, be it a conversation in a pub, a job interview, or a classroom discussion.

However, this view is akin to assessing driving by having a candidate drive alone on a closed course, and then assuming that they will do equally well in rush hour traffic. This is clearly a daring assumption: while the basic skills needed for successful driving (such as accelerating, braking and steering) are called upon in both situations, driving in rush hour traffic goes beyond basic driving skills. It requires coordinating one's actions with others, predicting what they might do, reacting to their actions, and adapting to changing conditions.

While driving on a closed course is like speaking, driving in rush hour traffic is like talking: one needs to be able to speak in order to talk (though not in all circumstances as will be elaborated), but talking requires more than speaking. In addition to basic speaking skills, talking also requires understanding of the interlocutor, and adjusting of one's speaking to the interlocutor. Both are crucial, so they will be discussed in turn.

Understanding of the interlocutor involves real-time decoding of incoming language on a semantic level of word meaning, as well as on a pragmatic level of social action and interpersonal meaning. To be able to respond, interlocutors must understand the content of what is being said, but to be able to respond appropriately, they must also understand what action is being done and how the utterance frames the relationship between the interactants. All of this happens before the interlocutor has even finished speaking as otherwise there would simply not be enough time, given that assembling a response takes about 0.6 seconds but gaps between turns in English-language conversation are only about 0.2 seconds (Levinson & Torreira, 2015).

Knowing what social action is being performed is crucial for designing a response, as social actions tend to only have a limited range of typical responses. For example, in responding to the informal greeting 'How is it going?', a response like 'Not bad, you?' is quite typical because it consists of a second pair part commonly associated with this type of greeting. However, an account of one's health status ('How is it going?' –'I've had really bad hay fever recently.') is a-typical, as it recasts the social action of greeting as an inquiry after well-being. Finally, a response like 'I'm going by train.' indicates a semantic lack of understanding and would most likely be taken as a mis-hearing, which may be repaired or ignored.

In addition to responding to a social action with a type-fitting action, utterances must be recipient-designed. In the CA literature (e.g., Drew, 2013), recipient design refers mostly to taking into account shared knowledge between interlocutors. For example, when a speaker talks about their partner, Mike, to a stranger who can be assumed not to know Mike, they are likely to refer to Mike as 'my partner'. However, when they talk to a close friend who knows Mike well, they are likely to refer to him as 'Mike'; in fact, referring to him as 'my partner' would be decidedly odd.

Talk can be specifically designed for a recipient in other ways as well. If we revisit the discussion on pragmatics theories in Section 2.1, Brown and Levinson's (1987) seminal work on the social context of interaction has focused on social factors that impact how people talk to each other. Here I will briefly

review the three contextual variables that Brown and Levinson (1987) theorized as influencing the politeness level of utterances: power, social distance, and rank of imposition. Power refers to the degree that one interlocutor can exert control over the other one's behaviour, e.g., in a workplace situation, a manager would have power over an employee they supervise. Social distance is the degree of acquaintanceship or common membership in a social group between interlocutors, e.g., close friends have low social distance whereas strangers have high social distance. Finally, rank of imposition describes the 'cost' to the hearer in terms of money, time, effort, or social sanction for complying with the speaker. For example, asking someone for the time is a low imposition request, but asking them to go out of their way to help carry a heavy item is high imposition. It is worth remembering from Section 2.1 that these factors are viewed in this book as not deterministically governing how people talk to each other, which would otherwise be an exclusively rationalist and utilitarian perspective on interaction (see Section 2.1.5 on how different theoretical perspectives on interaction are reconciled for IC assessment). In addition to the three contextual variables in Brown and Levinson (1987), there are other interactional and contextual factors that come into play. For example, Curl and Drew (2008), taking an empiricist-interactional approach to interaction, showed that the perceived degree of entitlement for making a request in emergency calls strongly affects request formulation.

Finally, work in sociolinguistics has identified a number of other factors that impact talking, most famously encapsulated in Hymes's (1974) SPEAKING model, which takes into account (among others) the physical setting of the interaction, its overall tone, the channel through which it is conducted, and the cultural norms through which interlocutors make judgments about meanings.

Given the numerous additional factors that impact talking, as opposed to speaking, it is difficult to imagine that simply measuring speaking/LC would give test designers a strong basis for making inferences about the ability in talking/IC. However, this is really an empirical question. If strong speaking ability invariably leads to strong talking ability, it is sufficient to test LC and then simply infer IC. This may seem unlikely, and speaking will probably not account for all the variance in talking, but even if it just accounts for a large amount of variance, that may be sufficient for a test. However, if a dissociation between speaking and talking can be demonstrated, where you can be good at speaking but not talking and vice versa, talking would need to be tested separately to enable defensible inferences from test scores. The below two sections discuss each of the two situations separately.

2.3.4 Strong on speaking/LC but weak on talking/IC

It certainly appears that way as anecdotal evidence abounds about test-takers who do well in the testing situation but not so well in real-life interaction, e.g., in medical communication (Eggly et al., 1999; Hall et al., 2004). In fact, end-user complaints about apparent mismatches between candidates' scores on the Occupational English Test (OET) taken by medical professionals and those test-takers' real-world performance was a major impetus in the revision of the OET to include a stronger focus on role-appropriate interaction in the rating criteria (Pill, 2016). This divergence is likely due to tests being rated on language-focused criteria, which tend to privilege speaking/LC, and not criteria indigenous to users (Jacoby, 1998), which tend to privilege talking/IC. Sato (2014) supported this conclusion, showing that different aspects of speaking performance matter to naïve judges assessing a performance on their general impression of the test-taker as a skilled communicator as compared to trained raters assessing on language-focused criteria. This dissociation between the criteria used in language tests and real-world language requirements also likely accounts for the less than overwhelming confidence of students, academics, and employers in the predictive value of language tests for real-world performance (see Murray et al., 2014, for IELTS, and Malone and Montee, 2014, for TOEFL).

It appears that speaking ability as measured by such a test is not a good predictor of the ability to talk in the real world, which does not bode well for language tests. However, two counter-arguments to this line of reasoning could be adduced to defend the assessment of speaking/LC: first, it could be claimed that specific-purpose language tests like the OET not only require a general ability to interact but a role-specific ability to interact. To put it simply, not only do you have to talk well, you have to talk recognizably like a doctor, nurse, dentist, etc. A potential argument is that this ability is not required in general proficiency tests. Second, it could be argued that no test performance ever extrapolates perfectly to real-world performance, and so the apparent gap between speaking in a test and talking in the real world is simply an unavoidable gap between elicited performances in a controlled test environment and actual performances *in the wild*. This book argues that both points are akin to desperate rear-guard battles trying to stave off the inevitable loss.

First, role-specific interactional abilities are always required in any form of interpersonal interaction. There is no such thing as language use unbounded to social roles. Interlocutors are always speaking as a friend, colleague, supervisor, customer, partner, neighbour, student with all the social requirements of talking appropriately to that role. These requirements can be subtle: Bella (2014)

found that although both L1 and L2 speakers of Greek had control over the pragmalinguistic tools for performing refusals in role plays, they deployed their pragmalinguistic tools differingly depending on their proficiency levels. In other words, just because you have grammar and vocabulary that you can mobilize in speaking does not mean you can deploy it in conventional ways when you are required to talk from the perspective of a particular social role. This is more obvious in specific-purpose assessment, but no different in other language tests.

Second, while language use in tests is not the same as language use in the real world, strengthening the extrapolation inference should be the main mission of language testers. O'Sullivan (2019) laid out the problem very clearly, and the larger the gap between test performance and real-world performance (i.e., the weaker the extrapolation inference in Kane's, 2006 framework), the less value test scores have for end users because they do not enable the decisions end users need them for. Tests based on a universe of generalization (UG) that is mostly chosen to be practically measurable will not do well on extrapolation. This is precisely the issue with measuring speaking versus measuring talking; measure speaking through monologic measures has an advantage in terms of administration, standardization, and scoring, but such tests do not measure the skills associated with talking, such as designing talk for the recipient, responding, organizing talk sequentially, enacting befitting social roles and so on. Similarly, tests that simply use interaction to elicit ratable samples of language to be scored without any reference to interactional abilities, sell the talking construct short. Scores from these tests run precisely the risk of being mismatched with stakeholder impressions of ability, which can eventually bring down the whole language testing enterprise: why should stakeholders go through the trouble of obtaining test scores if these scores do not tell them what they want to know? This is a broader question than just speaking/LC versus talking/IC but because of the pervasiveness and everyday necessity of being able to talk, speaking versus talking is a particularly important aspect of this problem in language testing[1].

1 It should be noted that stakeholder dissatisfaction with test-takers' real-world communication ability can also be caused by other factors. There can be a mismatch between the language test constructs and the communication skills valued in real life. The cut-scores employed by stakeholders can be inapposite. It is also worth investigating if stakeholder grievances are only directed at L2 speakers as in many circumstances such as healthcare communication, L1 speakers who do not need to take language tests can be found lacking in communication skills as well.

2.3.5 Strong on talking/IC but weak on speaking/LC

It may seem counter-intuitive that a language user could be better at talking/LC than speaking/IC since so far talking has been portrayed as 'speaking plus': interactants need basic speaking skills, but to talk they need to be able to configure and deploy them within a particular physical, social and interactional context. However, it is possible to talk with little or even no language proficiency. Levinson (2006) described an interaction during field research with a deaf signer on Rossel Island, an island in the far southeast of Papua New Guinea. Levinson and his interlocutor shared no language and little background knowledge but successfully managed a storytelling. While this may seem like an extreme case, it is actually fairly mundane: anybody who has travelled to a country without speaking the local language knows that it is possible to communicate to some extent through pointing, miming, and gesturing. This is not sufficient for discussing abstract, complex topics, and perhaps not observable in assessment tasks that require strong LC to complete, but it demonstrates that talking in the sense of interacting is possible with very little speaking proficiency.

A case closer to the experience of applied linguists is the well-known case study of Wes (Schmidt, 1983). Wes is an L1 Japanese speaker who migrated to the US as an adult, and Schmidt collected tape recordings and field notes of Wes's use of English over several years. He found little and stagnating grammatical development, but rapid development and high levels of performance in spoken discourse. Wes had an active social life with many friends and managed everyday interactions in English successfully. While Schmidt described Wes's language ability in order to test Schuman's acculturation theory (1978), the disjunct he found between LC and IC is a good illustration of a highly interactionally competent language user with limited linguistic control of the language. Schmidt wrote:

> If language is seen as a means of initiating, maintaining, and regulating relationships and carrying on the business of living, then perhaps Wes is a good learner. If one views language as a system of elements and rules, with syntax playing a major role, then Wes is clearly a very poor learner. (p. 164)

Wes was good at talking but not so good at speaking, and it is probably safe to say that he would not have performed well on formal language tests where his lack of grammatical accuracy and limited vocabulary range would have likely been penalized.

To conclude, the above discussion offers theoretical and conceptual evidence that speaking and talking are two overlapping but still distinctive forms of

competence. This lends support to assessing talking/IC in its own right in testing settings so as to better gauge test-takers' ability to interact in real life. As Section 2.3 is largely theoretical in its treatment of the differences between LC and IC, the next section, Section 2.4, focuses on the practical concerns in operationalizing an IC/talking construct in language assessment settings.

2.4 Defining an IC construct: An operational discussion

Building on the theoretical discussion on differentiating talking/IC from speaking/LC in Section 2.3, this section further specifies the considerations in operationalizing an IC construct in assessment settings, drawing on empirical studies and evidence.

2.4.1 Are we measuring talking/IC or speaking/LC?

Having examined the difference and mismatch between IC and LC, the most logical question for assessment researchers to address next is to assess IC in its own right, independent of LC. A test that claims to assess talking/IC needs to not only prove that it is indeed assessing a talking/IC construct as defined, but it also needs to be able to demonstrate that its IC construct is sufficiently different from other related test constructs, such as constructs based on traditional conceptualizations of speaking/LC. Another way to phrase the question is whether there is sufficient evidence demonstrating IC as a distinctive construct vis-à-vis traditional LC constructs already measured in existing speaking tests. From a test validity perspective, a test that claims to measure talking/IC should assess talking/IC instead of speaking/LC, even if there is a high correlation between talking/IC and speaking/LC. This is because in this case talking is a primary indicator of IC while speaking is at best a correlational, secondary indicator of IC. Using secondary indicators to substitute primary indicators for convenience can threaten the validity of a test and generate negative washback effects in teaching and learning (Scriven, 1987).

Though a strong argument can be made in regard to using primary instead of secondary indicators for the validity of the test, due to commercial and practical concerns, testing companies and relevant stakeholders are unlikely to abandon existing speaking tests in favour of IC tests unless the latter can be shown to cover variance not already covered by the former. If an IC construct claims to cover 'talking' variance not already covered by existing tests that measure 'speaking' or general proficiency, it needs to be able to demonstrate it in a statistical sense. Having this goal in mind and then looking at existing research in L2 IC, the

issue of differentiating between IC and LC quickly becomes a chicken-and-egg question.

Longitudinal studies on L2 IC development document L2 speakers' changing interactional methods or patterns but it is difficult to tease apart how much of those changes are attributable to an increase in IC or an increase in general proficiency (Hellermann, 2007, 2008; Pekarek Doehler & Berger, 2016). Cross-sectional studies, on the other hand, start with pre-grouping L2 speakers by proficiency and investigate if speakers from different proficiency levels mobilize different methods in implementing the same social action (Al-Gahtani & Roever, 2018; Pekarek Doehler & Pochon-Berger, 2011; Pekarek Doehler, 2018). Their grouping criteria are either proficiency frameworks based largely on the speaking construct (e.g., CEFR) or researchers' intuition. The study design of cross-sectional IC research similarly cannot definitively answer the question of whether differences in IC are caused by pre-existing differences in LC/proficiency.

The few IC assessment studies that have been conducted so far have also mostly adopted cross-sectional designs (Galaczi, 2013; Ikeda, 2021; Youn, 2015). Though they provided statistical evidence that L2 speakers did differ on researchers' IC measures, we cannot know for sure if proficiency or 'speaking variance' can already account for such IC differences. Another related issue is the IC rating scales developed can unintentionally conflate IC with LC. Both Youn (2015) and Ikeda (2017) recruited test-takers based on proficiency measures, grouped them by proficiency, inspected their test performance data using CA, and extracted group-specific interactional features for the steps in their IC rating scales. Although this approach can detect differences in test-taker performance, a counterargument is that the group-specific IC differences captured in their IC rating scales can be attributed to pre-existing group differences in proficiency/LC due to the grouping method employed, instead of substantive differences in IC. This is noticeable when we look at the correlations between IC scores and LC scores. In Youn (2013) the correlation between her pre-grouping proficiency measure and test-takers' IC test score was 0.9 ($p<0.01$) while in Ikeda (2017), it was 0.8 ($p<0.01$). If a proficiency/LC test can already effectively differentiate test-takers' IC, why is there a need for developing a new IC test?

The answer to this question is complex and is related to issues such as how the test construct is specified and how the rating scale is developed. The influence of LC needs to be controlled in the development of the IC rating scale to ensure the ensuing IC test construct is not contaminated by unwanted influence from LC. The definition of IC explicates the ability to orient appropriately to interlocutors' previous utterances and produce sequentially meaningful and relevant responses. However, such a theorization can prove challenging to operationalize when we

consider the role of LC in this definition of IC. The connection between IC and LC is frequently noted in IC literature when researchers discuss the relationship between L2 speakers' evolving IC and their developing linguistic repertoires (Pekarek Doehler & Pochon-Berger, 2015), L2 speakers' ability to apply grammar for interactional purposes (Pekarek Doehler & Berger, 2016), and L2 speakers' skills in monitoring their interlocutors' linguistic structures (Hellermann, 2009). Questions can be raised as to whether IC is actually tapping L2 speakers' lexical complexity, grammatical competence, and working memory capacity, which are all measures of LC. IC assessment researchers, therefore, need to consider in their rating scales if lexicon, grammar, and other measures of LC are to be included, and if they are, how these LC measures are interpreted through the lens of social action and are specified defensibly as serving for the purpose of interaction. For example, an IC test construct or rating scale should not reward a test-taker simply because of the emergence of a certain linguistic form (i.e., 'I was wondering' as a mitigator). It is only when such linguistic forms are demonstrated in the speaker's sequential organization as relevant and facilitative that they should be considered relevant to the IC construct. Therefore, strength in lexicon, grammar and other LC measures is not evidence of strong IC unless it can be proven to have fulfilled interactional purposes. These issues will be further explored in Section 2.6 on the development of IC rating scales but for now, let us have a closer look at some existing studies that have explored the relationships between IC and LC to better understand how differences between IC and LC manifest both qualitatively and quantitatively.

2.4.2 Separating IC from LC

The previous section discusses the need to separate IC and LC for assessment purposes, despite the interrelated relationship between the two constructs. In this section, we will look at how findings from previous research have generated both quantitative and qualitative evidence for the disconnection of IC from LC.

The first study to be examined is Lee and Hellermann (2014), which offers qualitative evidence on the separation between IC and LC using cross-sectional data. One of the excerpts in Lee and Hellermann (2014) focused on how Larissa, a low-proficiency speaker, demonstrated the capacity to launch a storytelling sequence despite her lack of linguistic resources. In the excerpt, Larissa from Russia was conversing with another low-proficiency speaker Jamie from Mexico, and before the turn that launched the story-telling sequence in focus, a sequence on Thanksgiving was completed with an agreement token by Larissa. After a 2.5-second gap, Larissa self-selected, and launched her story-telling sequence with

'mm husband #uh::# call my uncle.' (p. 772). Lee and Hellermann (2014) argued that although Larissa's story-telling announcement turn lacked explicit time reference devices such as 'yesterday', which are common in high-proficiency L2 speakers' talk, she still designed her turn to adumbrate a forthcoming story. Accompanying this turn, Larissa employed body language mimicking a person making a call, suggesting that her husband was calling her uncle to talk about the American tradition of having turkeys on Thanksgiving. It is also worth noting that in this turn Larissa evoked family categories 'husband' and 'uncle', which, combined with other linguistic and paralinguistic resources in other turns, paved the way for inferences such as 'husband is living with Larissa in the USA', 'husband is calling Larissa's uncle to tell the story of American Thanksgiving' and 'uncle is most likely still living in Russia'. Therefore, despite Larissa's linguistic inadequacies, she was still interactionally competent in securing a story-telling sequence through the mobilization of turn design, the employment of body language, and the evocation of social categories (e.g., 'husband' and 'uncle'). Regrettably, these markers of interactional resourcefulness are not covered in any existing speaking rubric.

Different from Lee and Hellermann (2014), the qualitative evidence in the second study, Jenks and Brandt (2013), takes place via CMC, which is the interactional platform the project in this book adopts. Although there is a paucity of research looking into the relationship between LC and IC in the CMC environment, there is similar emergent evidence suggesting the separability of talking/IC as a construct from the traditional speaking/LC constructs. Despite interactants' limitations in L2 LC, they can orient to the interactional norms of CMC and use a range of interactional methods to supplement their lack of L2 vocabulary and grammar. Jenks and Brandt (2013) did not have IC assessment as its main research question, but their data can be re-analysed to shed light on the difference between LC and IC. Situated in a voice-based online chat room, they reported observations on verbal alignment between L2 speakers of English. Here an excerpt is adapted to showcase how L2 speakers display IC despite their linguistic limitations. Excerpt 3 is taken from a conversation between Jin and Samir who were already in the chat room and Velson was the newcomer.

Excerpt 3 *1a – Skypecast, 09:47 – 10:00*

```
001  Jin     horror mo ↓ vie ↑
002          (0.8)
003  Samir   horror movie?
004          (0.9)
005  Jin     yeah sca↑ry↓  y'know horror movi:e
006          (3.0)
007  Samir   o ↑ kay
008  Velson  talk about mo ↓ vie ↑
009          (1.0)
010  Jin     >yeah yeah > we are talking about
011          movie
012          (0.6)
```

Note. data reproduced from Jenks and Brandt (2013) p238. Original line numbering is preserved

It is observable that Velson signals his willingness to participate in the discussion on horror movies between Jin and Samir in 008. Velson's 008 is not formatted as how it is usually done in English F2FC, probably due to a mixture of his lack of English LC and the medium of online voice chat. However, his first-pair part question format and rising intonation secure him a turn in an established conversation, and Jin orients to 008 as a question and provides an answer. Velson's successful attempt at transferring speakership to himself suggests that L2 speakers are not necessarily floundering simply because of their linguistic limitations nor are they struggling interactionally with the CMC platforms. This reanalysis suggests that there is some form of IC at work that guides L2 speakers through uncharted communicative territory, despite their limited LC.

Different from the qualitative evidence in Lee and Hellermann (2014) and Jenks and Brandt (2013), Ockey et al. (2015) is a testing study that compared IC measures with proficiency measures. The researchers examined 222 Japanese university students' TOEFL iBT performances in comparison with their performances on three language tasks (group discussion, picture description and oral presentation). From the account in Ockey et al. (2015), it seems that the rating scale and criteria used for the three language tasks were not derived from pre-existing LC differences based on test-takers' TOEFL scores, which is a different approach from the ones adopted in Youn (2013) and Ikeda (2017).

One component score for the three language tasks is IC, which is defined as 'participation and smoothness of interaction (e.g., turn-taking, responding to others, asking questions, introducing new gambits, paraphrasing, and hedging).' (Ockey et al., 2015, p. 46). Pearson correlations revealed high correlations between test-takers' TOEFL scores and their component scores on pronunciation, fluency and vocabulary in the three tasks (between 0.71 and 0.74, $p<0.05$). However, only a moderate correlation was achieved between TOEFL scores and their IC score ($r=0.63$, $p<0.05$). These findings offer evidence that TOEFL, a psycholinguistically-grounded speaking test, can provide a good prediction of real-life performances in terms of pronunciation and fluency but a weaker prediction when it comes to IC. It also shows that the asynchronous monologic speaking tasks employed by TOEFL can barely cover basic IC indicators such as turn-turning or hedging. It is speculated that higher-order talking/IC indicators such as action formation and social role enactment are even harder to predict from existing speaking constructs and rubrics, but such speculations are only tentative until corroborated by empirical evidence.

Finally, combining CA discourse analysis and statistical analysis, Roever and Ikeda (2021) is a recent mix-method study in IC assessment that looked at to what extent TOEFL speaking scores could predict test-takers' competence in an IC test. Although their correlation between TOEFL scores and IC test scores was moderate, with r ranging from 0.57–0.76 depending on test-taker cohorts, CA analysis of different test-taker performances from the same TOEFL score level showed great variation in test-taker ability at mobilizing interactional resources. This suggests that although IC and LC are overlapping constructs and LC scores can to some extent predict IC scores, IC tests cover variance unaccounted for by traditional LC measures, hence lending support to the argument that IC needs to be assessed in its own right.

In summary, Sections 2.4.1 and 2.4.2 shed light on the complex, interrelated relationship between LC and IC. In order to justify the assessment of IC from an operational perspective, IC testing needs to show that it is measuring a competence that is sufficiently different from LC. Previous research such as Youn (2013) and Ikeda (2017) has shown high correlations between IC and LC when the differentiation of IC abilities is to some extent based on pre-existing differences in LC. The correlation becomes weaker when the rating criteria of IC are unrelated to the ones used for LC, as observed in Ockey et al. (2015). Qualitative research such as Lee and Hellermann (2014) and Jenks and Brandt (2013) has also revealed the existence of IC indicators that differ more noticeably from LC indicators; the incorporation of such IC indicators can increase the value-add of IC assessment. Building on this line of argument, the following

sections will further explore potential IC indicators that index abilities that are unique to IC and unaccounted for in existing LC assessment.

2.4.3 Going beyond the mechanics of interaction: Hymes and Goffman revisited

Having examined the relationship between LC and IC in previous sections, this section explores what markers of interaction, or IC indicators, can be included in an IC construct to ensure that the IC construct is sufficiently different from an LC construct, making IC assessment a worthwhile undertaking. So far, the IC markers examined in L2 IC research are mostly sequential markers that index speakers' ability at progressing and preserving the temporal organization of talk, which May et al. (2020) labelled as the 'mechanics of interaction' (p. 3). Drawing on findings from CA, developmental IC studies have offered a wide array of potential mechanic and sequential IC markers such as progressivity (Balaman & Sert, 2017), alignment (Dings, 2014), post-expansion (Greer, 2016), dispreference structure (Al-Gahtani & Roever, 2012), and recipient design (Al-Gahtani & Roever, 2018). In IC assessment studies sequential markers such as turn organization, back-channelling, response tokens, and adjacency pairs have also been examined (Ikeda, 2017; Ross, 2018; Youn, 2013).

Although there is little restriction on what IC marker/feature to investigate from a research perspective, apart from the requirement that the intended markers/features need to be observable in performance data, the choice of IC markers in an operational language test requires stricter justification due to logistic and operational considerations. The markers selected should be consistently observable, ratable, and scalable in test-taker performances. They should be those aspects of human interaction that make a speaker like Wes (Schmidt, 1983) interactionally competent despite their linguistic restrictions. In terms of the process of scoring test performances, IC markers and the rating scales that explicate the markers are what translate test-taker performances into numeric scores. A poor choice of markers can potentially mislead raters, directing their attention to aspects of interaction not related to the IC construct. IC markers also carry the responsibility of construct validation as it is these markers that cover the added variance that existing speaking tests might fail to capture. Finally, due to the constraints of assessment, only a limited number of markers can be selected and incorporated into assessment rubrics. Therefore, the privileging of, for example, *topic management* as a marker over *turn-taking*, needs to have empirical grounding. Test designers should hence select markers that cover greater IC variance and that make a more tangible impact on the performance of talking/interaction.

Considering the mechanistic, sequential focus in existing IC assessment research, it might be fruitful to explore what other potential IC indicators exist that can offer a more comprehensive depiction of L2 IC. Harding (2021) suggested that for the field of language testing to move forward, language testers need to venture outside theoretical frameworks in the testing literature and explore concepts and theories in related fields such as sociology. Before we delve into sociological concepts, let us revisit a concept that has profoundly influenced applied linguistics and language assessment: Dell Hymes's communicative competence.

In response to Chomskyean's dichotomous view on competence and performance and his focus on linguistically perfect speech produced by ideal speakers, Hymes put forward the argument that language acquisition is not just about grammar but also about appropriateness. The original Hymes-ean conceptualization of communicative competence was influenced by works from Goffman (1956, 1963, 1964) and emphasized not only appropriate language in its broader sociocultural context, but also the interrelated 'attitudes, values, and motivations concerning language, its features and uses' (Hymes, 1972, p. 277). Hymes (1972) further argued: 'it is especially important not to separate cognitive from affective and volitive factors, so far as the impact of theory on educational practice is concerned' (Hymes, 1972, p. 283). It is clear that in Hymes's original conceptualization of communicative competence, effective language use was not solely a cognitive exercise as emphasized by psycholinguistic models of speaking. Instead, it was a multilayer process that not only draws on a speaker's cognitive abilities such as lexical range or grammar knowledge but also their attitudes and emotions towards language use. Further to this, building on concepts from Goffman (1967), Hymes went on to posit that 'capacities in interaction such as courage, gameness, gallantry, composure, presence of mind, dignity, stage confidence' should also be included in our understanding of a person's ability in language use/communicative competence (Hymes, 1972, p. 283). This extends the scope of communicative competence to include aspects of the speaker's personality, moral character, emotional intelligence, reasoning skills and other non-cognitive factors. Though Hymes (1972) did not prescribe a list of constituents of communicative competence, his explication of what he called communicative competence goes beyond the mechanistic IC indicators that IC researchers have investigated so far. If we want an IC construct that encompasses a wider range of abilities that characterize real-life language use and its effects, we need to move beyond the mechanistic, sequential tradition of IC.

2.4.4 Emotional, logical and moral IC markers

The previous section, Section 2.4.3, revisits the original Hymes-ean model of communicative competence, which suggests a wider range of communicative abilities for language use in real life, such as speakers' affective stance, reasoning skills and moral qualities. These abilities have been discussed in various forms in CA and ethnomethodology. Considering the empiricist-interactional approach adopted in this book (see Section 2.1.4 and 2.1.5), here a brief review is offered on the affective, reasoning, and moral IC markers to enrich our understanding of these dimensions of interaction from the CA perspective.

CA literature has discussed a speaker's affective stance in various forms such as affiliation (Lindström & Sorjonen, 2013), disaffiliation (Steensig, 2019), empathy (Stivers, 2008), frame (Kim, 2013) and emotion (Ruusuvuori, 2013). Here I will focus on *dis/affiliation* because they are closely tied to the concepts of *dis/preferred structures* in CA and the politeness theory concept of *face-threatening acts*, which have featured greatly in second language pragmatics research and have been investigated in IC assessment (Youn, 2015).

Dis/affiliation first appeared in CA literature in Heritage (1984) where the author contended that affiliative actions promote and support social solidarity, whereas disaffiliative actions threaten and damage it. Dis/affiliation is related to dis/preferred structure in that preferred structures are usually affiliative, whereas dispreferred structures are routinely disaffiliative (Heritage, 1984). The use of the term dis/affiliation in CA literature, however, goes beyond the sequence of turns, which traditionally has been the focus of research that examines the organization of preference. Compared with preference structures, dis/affiliation concerns more general features of stances, social actions, and social relations (Pomerantz & Heritage, 2013; Steensig & Drew, 2008). The resources for conducting affiliative and disaffliative moves include a wide range of lexical, grammatical, phonetic, and prosodic devices (Lindström & Sorjonen, 2013). The use of prosodic features to mark dis/affiliation has received growing attention in CA research (Couper-Kuhlen, 2012; Walker, 2013) but has been under-researched in IC assessment research. In fact, although much of our interaction relies on the concurrent prosodic features in our talk, prosody, as a potential indicator of test-takers' general IC, has not been systematically investigated in existing IC testing studies.

Compared with *dis/affiliation*, which has predominantly received interest in CA, *reasoning* as a concept has been formulated in the early days of ethnomethodology. Garfinkel (1967) proposed that common-sense reasoning undergirds social actions, making social actors' social actions recognizable to

each other. Heritage (1988) argued that both ethnomethodology and CA are concerned with two forms of reasoning. The first one is the normative, taken-for-granted reasoning that creates and sustains social actions. The second one is the explicit actions of providing accounts, explanations, and reasons. The first form of reasoning is indexing any form of social interaction whereas the second one is required when expected social behaviours are not forthcoming. For example, when an invitation is issued, the invitee needs to proffer an explanation for not being able to attend. The explanation itself relies on the second form of reasoning but the provision of the explanation relies on the tacit, taken-for-granted, common-sense first form of reasoning of how explanations are conventionally offered.

Similar to *reasoning*, *morality* as a concept in ethnomethodology has been conceived at both the discourse-internal and discourse-external levels. At the discourse-internal level, morality is defined as proto-morality (Bergmann, 1998), moral order 1 (Turnbull & Carpendale, 2001), and the moral order *of* interaction (Turowetz & Maynard, 2010). In this present discussion, the term proto-morality will be used. Proto-morality relies on the assumptions that first, social actors have the ability to choose amongst a range of courses of action. We can choose to be responsible or irresponsible, respectful or disrespectful, conditionally relevant or conditionally irrelevant. Second, this ability to choose is extended to other social actors whom we consider as having the same ability, which embodies the principle of the reciprocity of perspectives (Schutz, 1962). Since we consider our interactants as having the same ability to be responsible for their decisions, we have the right to make moral judgements of their choices. Because perspectives are reciprocated, we, in the process of morally assessing others, reciprocally invite others to make moral assessments of ourselves. This moves us to discourse-external morality, which is the explicit moral qualities we project of ourselves and ascribe to others. This is termed moral order 2 in Turnbull and Carpendale (2001) and the moral order *in* interaction (Turowetz & Maynard, 2010). Bergmann (1998) argued that this discourse-external morality is always culture-specific since different cultures form different understandings of moral qualities. Such moral qualities can be courage, gameness, gallantry, composure, and dignity, which are the personality features enumerated by Goffman (1967) and considered to be integral elements of communicative competence by Hymes (1972). The cultural specificity of moral qualities requires that moral claims and moral judgements be oriented to and interpreted in the specific sociocultural context where moral interaction takes place.

2.4.5 Aristotelian artistic proofs: Pathos, logos, and ethos

The previous section 2.4.4 offers a brief review of IC markers that index a speaker's ability to manage the emotional, logical, and moral aspects of interaction from a CA perspective. Apart from being mentioned in the Hymes-ean model of communication competence (see Section 2.4.3), the selection of and discussion on these three particular aspects of interaction might seem haphazard, considering the potentially unlimited list of indicators of IC. If we, however, venture out of theories familiar to applied linguists (Harding, 2021) and position IC assessment in more holistic models of interpersonal communication, we notice that affect, logic, and reason have been some of the most frequently mentioned dimensions of human interaction. One such holistic model of interpersonal communication is the tripartite model of persuasive rhetoric put forward by Aristotle.

In *On Rhetoric* (Aristotle, trans, 2007), Aristotle defined pathos, logos, and ethos as three modes of persuasion, or as he called them, artistic proofs. Aristotle initially conceived the three modes for orators delivering speeches to a public audience. The term artistic proofs implies that these three methods are means of persuasion, which are not evidence of the speaker's own quality or moral character. Pathos is persuasion through awakening emotions in the audience, logos is persuasion through revealing the truth of the argument, and ethos is persuasion through showing the speaker being worthy of credence. Modern rhetoricians have extended Aristotle's tripartite model and broadened the definitions of pathos, logos, and ethos, arguing that effective communication requires the connection with the audience on an emotional level, the presentation of a logical argument, and the demonstration of the moral character of the speaker (Aristotle, trans, 2007). These three criteria have been used as assessment criteria in academic fields such as rhetorical criticism (Afary, personal communication).

At a surface level, Aristotle's modes of persuasion might seem unrelated to the research questions commonly addressed in the field of applied linguistics and disjointed to the discussion on the conceptualization and assessment of IC. However, persuasion is a form of communication and successful persuasion requires effective language use. Coming from the perspective of IC, even though not every interactional setting requires artistic persuasion, a speaker still needs to engage with their interlocutor on the emotional, logical, and moral levels to make their interaction effective and appropriate, which is instrumental to successful interaction. Whether it is Aristotle's modes of persuasion or Hymes's definition of communicative competence, both models point to non-mechanistic dimensions of interaction and progress into dimensions of interaction pertaining to the emotional, the logical, and the moral. Though IC has already broadened

our traditional psycholinguistic understanding of speaking and has directed our attention to the sequential and co-constructed nature of talk, if we would like IC to better represent the reality and richness of interpersonal interaction, it will be conducive to examining if ability indicators from other more holistic models of interpersonal interaction can be incorporated into the theorization of IC.

2.4.6 Membership categorization analysis: Categorial IC markers

Finally, after discussing the emotional, the logical, and the moral dimensions of interaction, there is the categorial dimension of interaction, which similarly, has received little attention in IC assessment. The concepts 'category' and 'categorial' derive from Membership Categorization Analysis (MCA) and will be explained in this section.

As discussed in Section 2.3, there is no language use unbounded to social roles and this is true for language test-takers taking a language test. In addition to talking as a *test-taker*, in English for academic purposes (EAP) assessment a test-taker talks as a *student* or a *classmate* while in OET contexts a test-taker talks as a *doctor*, a *nurse* or a *physiotherapist*. The ability to perform particular social role(s), however, is largely ignored by existing language tests, with OET being a rare example. If the ability to enact and orient to social roles is to be incorporated in the conceptualization of IC, the analytic power of both CA and MCA should be combined to further our understanding of social role enactment as an IC marker. In this book I coined the term Sequential-Categorial Analysis (Dai, 2023b), which describes the methodology of concurrently conducting CA and MCA to understand the temporal and social aspects of interaction. At a more abstract level, I conceptualize the relationship between sequence and categorization as synonymous to the one between time and space, with change in one affecting the other. Interactants draw on both sequential and categorial resources in language to build a four-dimensional lifeworld where social actions take place. Sequential resources – such as turn-taking and preference organization – display interactants' understanding of the temporal nature of interaction. Categorial resources – such as membership categories and predicates, which will be introduced shortly – demonstrate interactants' ability to enact social roles, identities, physical objects, entities, interpersonal relationships, institutions and broader social organizations. These categorial resources are the building blocks for the three-dimensional spatial world, which, coupled with the fourth dimension time enacted in sequence, becomes the lifeworld where interaction takes place and where IC is observed. In order to appreciate how categorial

resources are marshalled by interactants to create a spatial understanding of the social world, recourse to MCA is necessitated.

MCA, originally developed by Harvey Sacks, is the study of how social members use categories (e.g., doctor, mother) to demonstrate their members' knowledge of the categorial dimension of interaction: namely what a speaker says or does at specific times depends on who they are and to whom they are talking. Though MCA developed rather slowly while CA flourished following Sack's seminal work (Sacks, 1992), we are now seeing a renaissance of interest in MCA (Stokoe, 2012), with an edited book on advances in MCA (Fitzgerald & Housley, 2015) and a special MCA section in *Journal of Pragmatics* (Fitzgerald et al., 2017). This shows that CA analysts have acknowledged that apart from the many sequential concerns of talk, there is also the categorial aspect of talk that is worthy of investigation. How speakers evoke categories such as *doctors*, *nurses,* and *students* index their members' knowledge of social roles in their society and culture. L2 developmental studies are yet to fully utilize the analytic tools and insight from MCA, though Lee and Hellermann (2014) offered some nascent findings in this regard (see the discussion in Section 2.4.2). MCA also has not made inroads into L2 IC assessment yet as IC rubrics so far still focus solely on sequential markers such as turn-taking, topic development, adjacency pairs, and back-channelling without situating such markers in test-takers' categorial knowledge (Dai & Davey, 2023; Galaczi, 2013; Ikeda, 2017; Youn, 2015).

When conducting MCA studies, researchers need to transcribe speech in a turn-by-turn fashion similar to the practice in CA studies. The difference between MCA and CA is that, apart from the sequential concerns in terms of how turns are generated, MCA also takes up a categorial focus on how speakers construct their knowledge of categories and culture through turns (Dai & Davey, 2023). The concept *category* is crucial to the understanding of MCA. Sacks (1972) defined categories as groups composed of members. The category 'student' can contain an infinite number of members who consider themselves to belong to this category and the interaction itself between the interactants will further specify what the category 'student' means *in* the interaction, *to* the interactants. A less technical term for category is social role, which I have already used before and will be using interchangeably with category.

Categories in interaction invoke knowledge about what is permissible, conventional, and expected for a particular category known to members of the same community. In some speech communities, there might be specific expectations as to how members from the 'student' category are supposed to behave vis-à-vis members from other categories, such as the category of 'teacher'.

These expectations can be about what social actions are permissible and how such social actions are realized linguistically and para-linguistically. Sacks (1974) used the term *activity* to describe the actions that are perceived to be paired with certain categories, such as 'mums' (category) are supposed to 'pick up their children' (activity) when the 'children' (category) 'cry' (activity). The activity 'picking up children' is paired with the category 'mother' and is expected of members from the category 'mother'. If a member supposedly belonging to a certain category does not perform the expected activities associated with that category, this member risks attracting sanctions and even condemnations from other members in that speech community. Schegloff wrote about the severity of a mismatch between category and activity by asserting that 'if an ostensible member of a category appears to contravene what is "known" about members of that category, then people do not revise that knowledge, but see the person as 'an exception', 'different', or even a defective member' (Schegloff, 2007a, p. 469). A lack of knowledge of social roles and their concurrent activities can create misunderstanding and threaten social solidarity. Considering the very real consequences of mismanaging social roles, the categorial aspect of interaction should be considered when developing a holistic assessment construct of IC. This highlights the advantage of Sequential-Categorial Analysis, which investigates both the sequential and categorial nature of interaction. Dai and Davey (2022, 2023) is a case study that demonstrates how Sequential-Categorial Analysis can uncover the systematic categorial methods speakers employ to talk social roles into existence.

In summary, Section 2.4.3 to Section 2.4.6 have surveyed some potential IC indicators that go beyond the sequential indicators that are traditionally researched in IC assessment, inspired by Hymes-ean communicative competence, Aristotelian modes of persuasion, and the oft-neglected categorial nature of interaction. It should be noted that the discussion of these potential markers is only tentative at this point as the selection of markers needs to match the specific assessment contexts and needs. A more methodical approach to the selection of IC indicators should be employed so as to ensure the chosen markers are crucial to interactional success in specific assessment tasks and contexts. These issues are explored in Section 2.6, which explicates a methodical process for the selection of IC indicators for assessment.

2.5 Designing IC test tasks

Having discussed the complexity involved in defining an IC test construct from the theoretical and operational perspectives, the following two sections,

Sections 2.5 and 2.6, focus on the processes involved in designing test tasks and rating materials. Both steps are crucial to the operationalization of an IC test since test tasks serve to elicit IC performance while rating materials translate IC performance to scores indicative of test-takers' IC.

The design of assessment tasks is an empirically motivated process, commonly initiated through a needs analysis (NA) on test-takers' language needs in the target language use (TLU) domain (Knoch & Macqueen, 2020). Section 2.5 first discusses how IC assessment can benefit from systematic NA before examining the practice of NA in test design and reviewing existing literature on NA.

2.5.1 From the target language domain to a test

The design of a language test starts with the delineation of the TLU domain. In terms of IC assessment, Galaczi and Taylor (2018, p. 227) offered a tree-shaped visual representation of IC which is helpful in visualising how the target domain frames the assessment of IC. Their IC tree grows into branches and leaves, with branches indexing general IC markers such as 'turn management' and 'interactive listening' and leaves indexing finer IC markers, such as 'maintaining turns' and 'pausing' under the branch of turn management. Going downwards, the authors specified the roots for the IC tree in three concentric circles: speech acts (or social actions in CA terminology), speech events, and speech situations. Though these contextual factors in the language use domain situate, nourish, and support the investigation of IC, they are rarely given the considerations they deserve. In terms of the first two circles, although previous L2 IC research has examined a range of social actions e.g., requests (Al-Gahtani & Roever, 2015), refusals (Al-Gahtani & Roever, 2018), storytelling (Watanabe, 2017; Waring, 2013) and disagreements (Pekarek Doehler & Pochon-Berger, 2011), the rationale for researchers' choices of social actions is rarely provided. Understandably for developmental IC studies, social actions merely serve as vehicles for analysts to observe differences in L2 speakers' IC so what social action is used to elicit their performance is not a primary concern. However, from a test design perspective, which social action in the TLU domain is included in or excluded from the test needs to be a carefully considered and empirically justifiable decision. For example, there need to be grounds for testing L2 speakers' ability to launch a request, instead of a disagreement or a complaint, although all social actions are fairly common and equally important to master in the TLU domain of everyday interaction.

When we move beyond speech acts and speech events in Galaczi and Taylor's (2018) IC tree, we encounter the largest circle, speech situations where

social actions take place and where broader sociocultural factors influence the launching of social actions. The authors' conceptualization of speech situations drew from Hymes (1974) where speech situations were decidedly sociocultural. This further complicates the picture for IC assessment. Cross-linguistic CA research has offered evidence that speakers adopt different interactional methods and sequentially structure their talk differently across languages and cultures when they conduct social actions (Golato, 2002; Huth, 2006). Developmental L2 IC studies have also noted that L2 speakers increasingly diversify their interactional methods to adapt to and align with routinized interactional patterns commonly found in the host culture (Cekaite, 2007; Pekarek Doehler & Pochon-Berger, 2015). The cultural specificity of social action implementation requires test designers to go one step further: they need to demonstrate the relevance of their chosen social actions to the particular L2 context and culture. In other words, the question now is not just choosing among requests, disagreements, or complaints for an unspecified L2 speaker population. It is to decide among these social actions for an IC test that measures the IC of speakers who speak a particular L2. For example, why should the ability to launch requests be singled out for speakers of a particular L2? Is it because requests in this L2 are especially frequent, complex, or difficult? Or do speakers of this L2 struggle with requests in particular due to some sequential, interactional, or sociocultural differences in how requests are formatted in the L2 or the L2 host community? Though every social action is worth investigating from a developmental perspective, due to the limited resources in testing settings, test designers need to define their target domain clearly and select social actions that are most germane to their respective L2 groups. Fortunately, there is a ready remedy for this problem: needs analysis (NA).

2.5.2 Task-based needs analysis

NA can assist L2 IC test designers in their depiction of the IC target domain and identify social actions that are most pertinent to or challenging for their specific test-taker population. A methodical NA serves this purpose and assists test designers with narrowing down the TLU domain to items in a test.

The design of a language test usually starts with understanding the needs of a particular learner group. NA in second language studies is a process where researchers collect information about a particular learner group's needs in the learning of the target language (Brown, 2009; Long, 2005a, 2005b; Swales, 2001). Such a process is crucial in the design and delivery of appropriate language teaching and assessment materials in order to better represent the real-world

communicative challenges faced by L2 speakers. When L2 speakers' needs centre on communication in real-world settings, findings from NA can effectively translate into meaningful communicative language tasks that serve pedagogical purposes (Long, 2013a, 2013b, 2015a, 2015b).

Though it is not uncommon to conduct NA on other aspects of language use (see Molle & Prior, 2008 for an NA on EAP writing), NA is predominantly associated with communicative language use in real-world settings (Long, 2013b). This makes NA frequently associated with effective task-based language teaching. In fact, most NA studies on communicative needs also detail how their NA findings can be translated into language tasks, making them task-based needs analyses (TBNAs) (Chaudron et al., 2005; Gilabert, 2005; Huh, 2006; Jasso-Aguilar, 2005; Lambert, 2010; Malicka et al., 2017; Martin & Adrada-Rafael, 2017; Mochizuki, 2017; Oliver et al., 2013; Youn, 2015). In particular, Long disambiguated his definition of tasks from three other definitions of tasks by reiterating that for him tasks are modelled on real-world communicative activities in which learners find themselves engaged (Long, 2016). Genuine language tasks have an aim in themselves because they are based on the analysis of actual language use, which is different from structure-trapping tasks that have an overriding concern in linguistic forms (Skehan, 1998).

The concept of tasks in Long's sense is highly compatible with the social actions that L2 IC researchers are interested in. Admittedly, how L2 IC interacts with L2 grammar (Pekarek Doehler, 2018) and how L2 IC relates to L2 speakers' LC still require further research. However, as discussed in Sections 2.3 and 2.4, IC goes beyond L2 grammar and traditional constructs of LC as IC requires L2 speakers to demonstrate the ability to jointly construct orderly social interaction and enact believable social roles (Gardner & Wagner, 2004, Roever & Dai, 2021). Such a competence is fundamentally functional and cannot be fully revealed by structure-trapping tasks such as discrete, decontextualized grammatical items. Only through genuine language tasks modelled on real-world activities can we observe language use in meaningful interaction and pinpoint L2 speakers' proceduralized knowledge of interaction. Therefore, a TBNA on L2 speakers' interactional needs can offer guidance for the design of authentic language tasks, which then serve as tools for the assessment of IC.

2.5.3 Triangulation in needs analysis

In order to design believable and useful language tasks, needs analysts generally start with developing a defensible methodology that best captures their target

learner group's communicative wants. One of the techniques commonly used in NA to enhance methodological validity is triangulation (Chaudron et al., 2005; Cowling, 2007; Dai, 2023a; Dai & Roever, 2019; Lambert, 2010; Long, 2013a; Rylander, 2011). Brown (2001) identified seven types of triangulation, which include triangulation of data sources, investigators, need analysis theories, procedures, disciplines, data gathering times and sites. Such an exhaustive approach is rarely implementable due to financial and time constraints. After surveying 39 NA studies on English for specific purposes in the last thirty years, Serafini et al. (2015) argued that a more actionable starting point is to consider how information sources and information elicitation methods can be triangulated.

There are many benefits in going beyond one information source when conducting an NA. Though it is not uncommon that second language NA studies rely solely on language learners as informants (Cabinda, 2013; Chaudron et al., 2005; Holme & Chalauisaeng, 2006), such an approach does not engage the perspectives of other relevant stakeholders. Even in-service learners, defined as learners that actively use the target language for communicative purposes (Long, 2005a; Serafini et al., 2015), cannot always clearly identify the sources of their interactional plights. Different perspectives, such as the ones from language teachers, young learners' parents and learners' host family members, can offer a more holistic picture of where learners fall short of their respective goals (Brown, 2009).

Triangulation of information elicitation methods is also a widely employed technique, considering that each method has its strengths and weaknesses. In Malicka et al. (2017) the researchers supplemented their main method interview with non-participant observation as a secondary method. Their study is concerned with the communicative needs of hotel receptionists, but the researchers noticed none of their interviewees mentioned small talk as a relevant task, despite its frequent occurrence in researchers' observation of receptionists' daily routine. Small talk is a necessary social lubricant and if mishandled, can have real repercussions for the rapport between receptionists and their customers. Such crucial information could have gone unnoticed had the study utilized only interviews as its elicitation method.

Within the field of IC assessment, Youn (2015) adopted triangulation and grounded the design of her IC test in an NA on the EAP domain. Her NA, later published as Youn (2018), triangulated the perspectives of programme administrators, instructors, and students, utilizing two data elicitation methods: interviews and questionnaires. Her data revealed a wide range of social

actions where EAP students struggle, which then fed into the two interactional tasks she designed. The two tasks in Youn (2015) required test-takers to role-play with a professor and a classmate and elicited social actions such as making a request for a recommendation letter and agreeing on a meeting time.

In summary, test designers should approach TBNA as a multi-stage, multi-angle process. The sources and methods selected should complement each other and jointly produce a balanced picture of learners' language needs.

2.5.4 Paucity of TBNA in L2 Chinese

Though NA is a process indispensable in any methodical attempt at the design of teaching and testing materials, a survey of recent NA literature shows that most studies are conducted on English as a foreign language (Lambert, 2010; Oliver et al., 2013; Park & Slater, 2014; Sawaki, 2017; Serafini et al., 2015) and usually for academic purposes or business settings.

Despite the exponential growth in the L2-Chinese speaker population (CCIS, 2019), Wang (2011) is one of the very few NA studies conducted on L2-Chinese speakers' learning needs. Wang (2011) is a commissioned project that combined NA with Chinese business textbook analysis. Using interviews and questionnaires as research methods and eliciting perspectives from both learners and Chinese business managers, Wang identified a list of top-priority abilities business Chinese learners are expected to develop. Subsequent textbook analysis revealed a mismatch between learners' wants and textbook materials. Topics such as answering job interview questions, though highly valued by Wang's informants, were poorly represented in the 44 business Chinese textbooks surveyed. Findings from Wang (2011) corroborate Long's argument that most commercial language teaching materials on the market are 'written, on the basis of textbook writers' intuitions, for all students and for no students in particular' (Long, 2016, p. 6). The paucity of NA on L2 Chinese, coupled with the growing demand from L2-Chinese speakers and the real-world relevance of IC teaching and assessment, points to the necessity of conducting an NA specifically investigating the interactional needs of L2-Chinese speakers. Considering the time, money and effort L2-Chinese speakers invest in mastering Chinese, it is regretful that no studies have examined what these speakers actually struggle with when they interact in Chinese. A TBNA of L2-Chinese speakers' interactional needs can not only inform learners, teachers, and other stakeholders of common communication pitfalls, but it can also serve as a starting point for effective L2 Chinese teaching and assessment with a focus on IC. Considering the lack of IC research and NA research on L2 Chinese, this book project chooses Chinese

as its target language, although the principles and methodologies of IC test development and validation explicated in this book are applicable to other languages as well.

2.6 Designing IC rating materials

Section 2.5 reviews existing literature on test task design and identifies TBNA as a starting point for the development of tasks that can be used for the assessment of L2-Chinese IC. In this section, Section 2.6, I will survey previous studies on the process of rating material development, which generates rubrics that embody the test construct and translate test-taker performance to IC test scores. As discussed in Sections 2.3 and 2.4, assessing L2 IC offers a sociolinguistic-interactional alternative to speaking assessment, which traditionally adopts a Leveltian psycholinguistic-individualist framework and focuses on LC speech markers such as accuracy and fluency (Roever & Kasper, 2018). Existing L2 IC studies have utilized findings in CA and the analytical policy of CA to identify IC markers pertinent to their respective test constructs and validate the constructs through CA inspections on test-takers' performance data (Ikeda, 2017; Youn, 2013). The value-add of CA to the scoring process, so far, has received much less attention. The various ways CA analysis can contribute to test design and validation will be further discussed in Chapter 3, but it is worth noting at this point that CA plays a crucial role in test construct specification and rating material development in this book.

2.6.1 IC rating materials development

When assessing speaking performance, the discussion on rating material development usually starts with the design of rating scales. The theorization of rating scale construction started in the 1990s and since then, different views and practices regarding scale development have emerged (Upshur & Turner, 1995; Turner & Upshur, 2002). Recent reviews on scale development have identified various practices such as basing the scales on the learning goals in L2 syllabi and on their extrapolations to real-world functional language use (Jamieson & Poonpon, 2013). Though different practices co-exist, literature in rating scale design has predominantly classified rubric construction endeavours binarily, according to whether they adopt an *a priori* intuitive approach or an empirical data-driven approach (Fulcher et al., 2011). The intuitive approach relies on scale developers' judgement in the construction of the criteria and descriptors of language proficiency in the scales (North & Schneider, 1998). Despite its

prevalent employment in L2 teaching and assessment, the intuitive approach has been criticized for the vagueness of its performance descriptors and its lack of attention to assessment contexts (Fulcher, 2003). The data-driven approach bases the scale on empirical data, which can either be performance data from test-takers or perception data from raters (Ducasse & Brown, 2009). It should be noted that this dichotomous classification oversimplifies the complex and sometimes iterative process of rating scale development. After conducting a systematic review of rubric development literature, Knoch et al. (2021) identified ten information sources that scale developers have drawn on to inform the construct of their scales, which are existing theories, TLU domain, standard frameworks, curriculums, expert intuition, existing scales, performance samples, rater input, task features, and score use context.

Within the field of IC assessment, researchers have predominantly favoured test-taker performance samples as an information source for scale development. One example is Youn (2015) where the researcher adopted an inductive approach in the design of her IC rating categories from CA inspection of test-takers' performance data at different proficiency levels. Kley (2019) is another study that used performance data to elaborate the description of the CA-informed assessment criterion *topic management*. The appeal of performance data to IC test developers is self-evident, considering CA's overriding focus on empirically substantiating claims with context-specific moment-by-moment interactional data. Despite IC rubrics' expected amalgamation with performance data, basing IC rubrics solely on test developers' CA analysis of performance data raises concerns in terms of whether the IC features identified by CA-trained test developers are actually the features considered crucial to interactional success by raters. To complement insight gleaned from performance data, test developers can draw on raters' perceptions of test-takers' performance in the construction of rating scales.

2.6.2 The rater perspective and indigenous criteria

Among the ten sources of information for scale construction in Knoch et al. (2021), the use of rater input has flourished in recent years (Ducasse & Brown, 2009; May, 2009; May et al., 2020; Pollitt & Murray, 1996; Sandlund & Sundqvist, 2019; Turner & Upshur, 1996). Collecting and analysing raters' perceptions of test-taker performance through think-aloud protocols, rater notes, and rater interviews can assist test developers with understanding what features in test-taker performance are salient to raters and considered crucial to fulfilling the tasks by raters. A recent study by May et al. (2020) showcases how test

examiners' perspectives can shape the development of IC teaching and testing materials. The researchers elicited 72 verbal reports from six trained examiners after having them view 12 paired interactions from test-takers. Verbal reports were thematically coded to provide empirical grounding for the design of IC checklists and feedback materials, which can be utilized by teachers and students in the acquisition and assessment of IC. One concern about the use of rater-perception data is that the rating categories raters focus on can be a-theoretical if the recruited raters do not possess the theoretical knowledge that applied linguists do (Brindley, 1998; Shohamy, 1996). The absence of theorization in raters' judgement can be mediated if scale developers adopt a mix-method approach through the combination of raters' perception with scale developers' theory-informed judgement in the construction of the scales. Roever and Dai (2021) argue that raters' perspectives play an oft-neglected but indispensable role in IC rating materials construction, where IC test developers constantly grapple with IC markers that pose challenges in terms of saliency to raters, scalability in rating scale design, and rater reliability in the psychometrically-driven large-scale assessment contexts.

One particular type of raters used in defining the test construct is called domain experts (DE), who are usually not linguistically trained (linguistic laypersons) but are knowledgeable of the interactional norms in the particular TLU domain in which the test is based (Elder & McNamara, 2016; Knoch & Macqueen, 2020; Pill, 2016). Domain experts' criteria of performance are termed their indigenous criteria (Jacoby & McNamara, 1999; Knoch & Macqueen, 2020; Pill, 2016). Drawing on the domain expert perspective, Pill (2016) developed two additional indigenous criteria, *clinician engagement* and *management of interaction* for the OET speaking test. The author arrived at both criteria by first obtaining medical educators' and clinical supervisors' judgments on simulated clinical performances from trainee healthcare professionals. He then conducted a thematic analysis of the data and generated the two new criteria, which were not covered by the previous OET rubric. These two criteria, similar to other criteria considered relevant by raters, can be argued to be a-theoretical from a linguistic perspective, especially when the DEs in Pill (2016) were clinicians instead of linguists or language educators. However, if we look closely at the scale descriptors of these two new indigenous criteria, we see that they are tapping dimensions of interaction that are commonly investigated in CA and MCA. Though Pill did not employ CA, it is evident that CA can be productively applied to the analysis of the *management of interaction* criterion to collect micro-level evidence of how interaction is co-constructed. This is a sequential IC marker that indexes speakers' ability to sustain an interaction. The criterion

clinical engagement, on the other hand, requires the ability to talk in a manner befitting the role of a doctor, a physiotherapist, or a nurse. In other words, medical professionals' language should evoke their respective socio-professional categories in relation to their interlocutors. Sequential CA, with its predominant concern with sequential properties, is restricted in its ability to collect such information. This suggests the necessary application of MCA, as introduced in Section 2.4.6, to explicate the categorial knowledge speakers need to possess to enact believable social roles. The foregoing analysis of the two indigenous criteria in Pill (2016) shows that indigenous criteria are not incompatible with existing theories of interaction as long as the researchers can theorize DEs' a-theoretical criteria with concepts in relevant theories.

Adopting the rater perspective in test construct specification, Sato and McNamara (2019) is another study that elicited indigenous criteria of general L2-English communicative ability from linguistic laypersons, who were L1 and L2 English speakers untrained in linguistics or language teaching. The recruitment of linguistic laypersons for their study is novel in the sense that L2-English speakers in real life are not always assessed or judged by language experts on their communicative ability. In fact, it is these linguistic laypersons, whether L1 or L2 English speakers, who interact most with L2-English speakers and form assessments of their communicative competence in everyday-life settings. Although Sato and McNamara (2019) did not use the term *DEs* but used *linguistic laypersons* instead to describe everyday-life members of society, here I propose the argument that *linguistic layperson* informants can be defined as *DEs* if the informants are purposefully selected and if the TLU domain is general/everyday language use. This proposition is in line with how Hymes (1972) conceived communicative competence in its general sense. Hymes considered communicative competence as an ability that is shared by normal members of a speech community. Normal or everyday-life members are able to use their innate judgement of what speech conduct is appropriate and feasible in the language use domain in which they frequently use the language and of which they have developed an innate knowledge. This argument was presented in Hymes (1972) as such:

> There is an important sense in which a normal member of a community has knowledge with respect to all these aspects of the communicative systems available to him. He will interpret or assess the conduct of others and himself in ways that reflect a knowledge of each (possible, feasible, appropriate), done (if so, how often). There is an important sense in which he would be said to have a capability with regard to each. (p. 282)

In view of this, in this book, the term *everyday-life DEs* is used to describe linguistic-layperson informants who are familiar with and have expertise in the everyday-life TLU domain and can serve as DEs to inform the relevant test construct.

2.6.3 Test-taker exemplars in IC rating

The previous section, Section 2.6.2, explicates the use of rater perspective and test performance samples in developing a test construct and focuses on a particular type of raters in the TLU domain: DEs. If the TLU domain is everyday-life language use and the purpose of testing is general proficiency testing, test developers can use everyday-life DEs who are linguistically untrained but nonetheless possess the lived experiences in the everyday-life TLU domain. Their indigenous criteria can be translated to assessment criteria in a rating scale.

Developing a rating scale to measure test-taker performance is a crucial process in rating material development. It is however not the only instrument that needs to be developed for effective rating. When using the rating scale to rate test-takers' speaking performance, raters need other rating aids to assist them with establishing the connection between the rating scale and the performances to be rated. The actual rating process is a complex one as raters need to be able to base their judgement on the assessment criteria, identify features in performance data that are relevant to the assessment criteria, and judge the quality of the performance on the specific features foregrounded by the rating scale. This process can be particularly challenging for IC assessment as raters might not agree on the features to be assessed, leading to concerns about rater reliability. Though it was not set up as an IC rating study, Walters (2007) highlighted the reliability challenges besetting IC test developers by investigating if CA-trained raters could reliably assess *the management of adjacency pairs*, one of the core IC features, in test-takers' interaction data. Apart from a rating rubric, the raters in Walters (2007) were given a rating sheet where they could transcribe test-takers' deployment of adjacency pairs in CA fashion to provide empirical evidence to justify their rating. An analysis of the rating results and raters' transcripts evinced that raters could identify and parse test-takers' adjacency pairs, but absolute inter-rater reliability was only 40 %. This suggests that raters were either assessing different facets of the IC construct or were unable to use the rating scale effectively. Findings from Walters (2007) underline the challenges in IC assessment in terms of rating and the need to develop rating aids for raters. Despite years of training in CA, raters in Walters (2007) still struggled with what to assess in relation to the CA-informed markers. In the realm of IC assessment,

more thought clearly needs to go into the design of training materials to standardize raters' understanding of IC markers in order to achieve the intra- and inter-rater reliability needed in large-scale assessment. One approach to address this challenge is the use of test-taker exemplar responses.

As a scoring aid for performance assessment, exemplar responses bridge the descriptors in rubrics and actual performance data from test-takers, making explicit to raters how the desired characteristics in a rating scale are reflected in empirical data. Despite its crucial role in rater training, exemplars have received far less attention in the rating literature compared to rubrics. Davis (2016) is a rare exception that sheds light on the relationship between the use of exemplars and rating accuracy. The author noted that more accurate raters used exemplars more often and they did so during scoring, whereas less accurate raters viewed exemplars fewer times and usually only did that before scoring. Findings from Davis (2016) support the argument that periodic review of exemplars can help align raters' interpretation of performance with the one envisaged by the test developer, a desirable rater training outcome that rubrics alone are unable to achieve. The benefits of exemplars can be more noticeable in IC assessment where IC markers are still at a nascent stage of being incorporated in large-scale language testing. Compared to well-researched psycholinguistic speaking markers such as fluency and accuracy, IC test developers can make greater use of explicit exemplars so that raters can precisely locate IC markers in test-taker performance.

Within the field of IC, Greer (2020) demonstrates how CA-transcribed exemplars can unpack differing IC performances. The author investigated student test-takers' interaction in a paired discussion format and its rating scheme included fluency, accuracy, complexity, and engagement. The raters in the study were familiar with the first three assessment criteria but were foreign to the rating category/criterion *engagement*, especially from the CA-for-IC lens. In order to explicate this IC criterion and standardize raters' interpretation of it, the author selected representative exemplars from four band levels, transcribed the exemplars in CA style, and provided descriptions to inform raters why a specific 'engagement' exemplar merited a particular band score, drawing on insight from the CA transcripts. Greer then used post-rating hermeneutic dialogue (Walters, 2007) to show that raters were able to make sense of the IC criterion *engagement* when they were provided with descriptive CA-informed exemplars. This study offers a promising direction to how CA-informed exemplars can assist raters with locating IC criteria in test-taker performance data and standardize raters' interpretation of differing performances on IC measures. This approach to rater training will also be taken up in this book project.

In summary, Section 2.6 situates the development of an IC test construct and rating scale in the mixed-method approach of combining the rater perspective with discourse analysis of test-taker performance. For this book project, the raters are further narrowed down to everyday-life DEs whose indigenous criteria of IC performance can serve as guidelines for rating scale and construct development. Apart from a discussion on rating scale construction, Section 2.6 highlights the necessity of developing target test-taker exemplars to assist raters in connecting test-taker performance with the rating scale. The use of CA in the generation of CA-informed exemplars can be particularly helpful to IC rating.

Chapter 3 Interpretive argument and research design

Chapter 2 offers a review of existing literature on IC assessment and positions this book in the area of assessing L2 IC for L2-Chinese speakers in the CMC context. The design and operationalization of a language test, however, involves test validation, which ensures the test is a valid instrument and the resulting test scores are meaningful and useful. Test validation in this book follows Kane's argument-based framework (Chapelle, 2020; Kane, 2006, 2013) and is separated into the interpretive argument in Chapter 3 and the validity argument in Chapter 7. The interpretive argument specifies the assumptions and inferences that need to be validated before the test can be trusted to be a valid instrument. The book is therefore designed to garner supporting evidence to substantiate the assumptions and inferences in the interpretive argument.

Chapter 3 consists of two sections. In Section 3.1 I draw on Kane's argument-based validation framework and proposes a validation process that examines the domain description inference, the evaluation inference, the generalization inference, the explanation inference, and the extrapolation inference. In order to specify the backings needed for each of the five inferences, Section 3.1 provides a detailed account of the interpretive argument for test validation, with the warrant and assumptions for each of the inferences explicated. The specification of the assumptions in turn shapes the design of the book, as the book is structured to collect the requisite backings to support the interpretive argument.

Having laid out the warrants and assumptions for the inferences in the validation framework in Section 3.1, Section 3.2 proposes three interrelated studies and explains how each study can gather evidence to support relevant assumptions in the interpretive argument. The three studies respectively focus on TBNA and test design (Chapter 4), test construct development and rating scale design (Chapter 5), and the main testing study (Chapter 6). After Chapters 4, 5 and 6 have reported the evidence generated by the three studies in this book in support of the inferences and assumptions, Chapter 7, the validity argument, critically reviews the evidence and assesses if the assumptions in the interpretive argument are sufficiently supported. If they are, the test can be considered a valid instrument to elicit IC performance and generate IC scores that are valid, meaningful and useful to end-users.

3.1 The inferences and assumptions in the interpretive argument

Building on Messick's (1989) unified model of validity, Kane's (2006) argument-based approach to validation has proven instructive to language testers as it offers a clear explication of how the inferences in test validation link to one another and produce a coherent argument for test validity. Since its inception, Kane's argument-based validation framework has been frequently adopted by language test designers (Chapelle, 2008; Chapelle et al., 2008; Ikeda, 2017; Knoch & Elder, 2013; Mendoza & Knoch, 2018; Roever et al., 2014; Trace, 2016; Youn, 2013). The inferences and assumptions to be investigated in this book are laid out in this section. It should be noted that the decision and consequence inferences (Knoch & Chapelle, 2017) are not examined in the validation process in this book because both inferences can only be investigated when the IC test is fully operational and in use. In terms of intended use, the test is designed to assess L2 Chinese speakers' ability to interact successfully when they migrate to China for study, work, and everyday life. The interpretive argument framework and the related inferences for this book are presented in Figure 1.

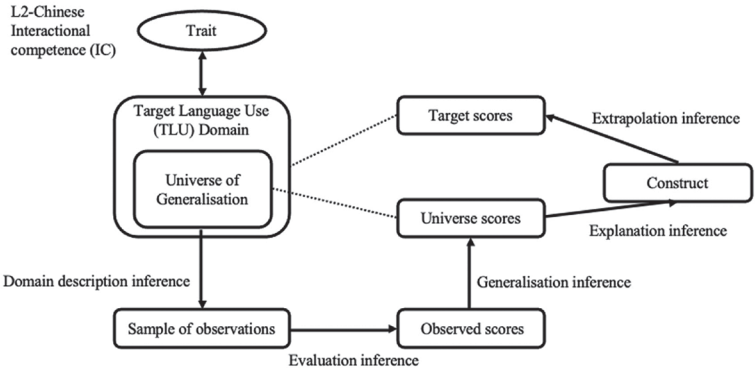


Figure 1 The validation framework for this book

3.1.1 The domain description inference

The domain description inference is the first inference in the interpretive argument in Figure 1, seeking backing for how a target language use (TLU) domain is described for assessment purposes (Chapelle et al., 2008). As discussed in Section 2.5.1, interpersonal interaction as a TLU domain can be very broad in

that it can encompass anything from texting friends about an upcoming concert, having small talk with colleagues at work, to presenting research findings at an academic conference. A more specific TLU domain needs to be established to narrow down the context for the IC test to be developed. If an IC test is measuring the IC of L2-Chinese international students' ability to communicate verbally in academic settings in universities where Chinese is the medium of instruction, we need to narrow down the TLU domain to a domain that covers the specific activities that L2-Chinese international students frequently encounter in tertiary education contexts where the mode of communication is spoken Chinese. This narrower domain has the advantage of having several parameters such as L2-Chinese, international students, university context and spoken language to circumscribe the domain under investigation, which makes identifying test tasks within the domain more manageable.

Despite the advantages of having a narrower domain in test design, the trade-off here is that too narrow a domain can limit the inferences one can draw from a language test. An even narrower domain than the one stated above could be, for example, one that focuses on the IC of L2-Chinese test-takers' ability to interact with L2-Chinese classmates in tutorial discussions in a third-year Chinese literature class at a university in China. This test is clearly easier to develop but the meaning of its score does not go beyond its immediate domain. We cannot be confident that the test scores can give us reliable information about test-takers' ability to communicate in other activities in academic settings, let alone their ability to communicate with friends or colleagues at work outside of university. Given the resources that go into test development, test developers working on large-scale standardized testing projects would be inclined to develop language tests that have a wider range of applicability. A test with a larger TLU domain will also generate more useful scores for stakeholders as the scores reflect test-takers' ability in a larger range of language use contexts, allowing stakeholders to make broader inferences in regard to test-takers' ability.

In view of this, I approach the delineation of the TLU domain for the IC test in this book in a broad-to-narrow approach. The broad-to-narrow approach means I refrain from imposing too many restrictions on the TLU domain at the start of the test design and instead, I define the TLU domain broadly as L2-Chinese interaction for L2-Chinese speakers who have a need to live, study and work in China. This domain is understandably very general and needs to be narrowed down to make a language test based on the domain operationalizable. The process of narrowing it down will be achieved through a task-based needs analysis (TBNA). The many advantages of a TBNA for test design have already been discussed in Section 2.5. Here I highlight in particular how a TBNA can

contribute to this book. Given that the current TLU domain at this stage is very broad, a TBNA can elicit informants' opinions on what to focus on in this domain and subsequently narrow the domain down to make test design feasible. Details on how I conduct the TBNA to refine the TLU domain can be found in Chapter 4.

Once a narrower, testable TLU domain is established, we can interrogate the domain description inference in the interpretive argument. The warrant for the domain description inference is that test-taker performance on the IC test can reveal test-taker IC in the L2-Chinese TLU domain. In order for this warrant to become legitimate, I have identified four assumptions, which are presented in Table 1. Now let us inspect each of the assumptions in turn and consider what type of evidence, or backing in Toulmin's (2003) terminology, is needed to make the assumptions justifiable.

Table 1 Warrant and assumptions for the domain description inference

Domain description inference	
Warrant	Test-taker performance on the IC test can reveal test-taker IC in the L2-Chinese TLU domain.
Domain assumption 1	Critical activities indexing IC can be identified in the TLU domain.
Domain assumption 2	IC assessment tasks can be developed that are based on the identified critical IC activities.
Domain assumption 3	IC tasks offer sound coverage of the TLU domain.
Domain assumption 4	The task scenarios and task delivery methods are valid and representative of activities in the TLU domain.

Domain assumption 1 states that the researcher can identify critical activities in the TLU domain that are reflective of test-takers' IC. The backing for this assumption can be gathered from a TBNA. Informants in the TBNA can report what they consider to be crucial IC abilities in the TLU domain, and the researcher can assess if their reports align with the definition of the TLU domain. As discussed in Section 2.5, a TBNA can assist the researcher with identifying what particular IC skills are considered crucial for interactional success in a specific target language setting. Since TBNAs focus on specific languages and contexts for language use, they can offer justification for the assessment of, for example, the social action of initiating refusals in L2-Chinese, if this social action

is exactly what L2-Chinese speakers find challenging. This addresses the concern that existing IC research does not always provide a rationale for the analyses of certain social actions over others when analysing the IC of a particular L2 speaking group, which is discussed in Section 2.5.1.

A TBNA can also define how IC is interpreted and assessed. IC can be defined as knowledge of interactional norms (Walters, 2007) or the ability to actually interact, whether in naturalistic settings or in controlled test settings such as discussion activities or role-plays. The informants for a TBNA can help test designers identify whether online IC performance or offline IC knowledge should be prioritized, which will have direct implications on the test methods chosen for the test. For offline IC knowledge test designers can choose multiple choice questions, discourse completion tasks or even writing tasks where test-takers are required to demonstrate their ability to interact in the written mode, such as text-messaging (Abe & Roever, 2019). If the results of the TBNA suggest a stronger focus on online IC performance, the researcher then needs to select test methods that are more interactive and that tap test-takers' online real-time IC behaviour. Suitable methods include timed speaking tasks such as leaving voice messages or synchronous role-plays.

Domain assumption 2 asserts that the IC tasks developed match the identified critical IC activities in the TLU domain. This assumption is pertinent to the nature of assessment, which in most cases can only elicit a slice of reality, instead of a fully replicated representation of reality. An example can be the informants in the TBNA report that L2-Chinese speakers find verbally refusing invitations from friends very challenging. In real life, such a social event can be a complex endeavour where the inviter may first suggest hosting a party when talking to a group of friends including the test-taker invitee. The test-taker invitee might have agreed on the spot but notices a few days later that the inviter's party time clashes with a prior commitment. The test-taker invitee might then try to bring up their inability to attend the party in an off-hand manner when talking to the inviter about a different topic on a different occasion. A standardized language test is usually unable to encapsulate in a test task the richness and complexity of such real-life interactional events, so abstraction and simplification are necessitated. The process of simplification will unavoidably decrease the authenticity of the test tasks and underrepresent the activities in the TLU domain, reducing the test construct and the inferences that can be drawn from the test. Test designers need to find a balance between authenticity and practicality so that the tasks are representative of real-world activities but are still logistically feasible to operationalize in testing settings. If we take the previous refusal activity as an example, a task that requires the test-taker to send a voice message to the inviter

about their inability to attend their party can still reasonably elicit the IC ability in the original real-life activity in the TLU domain. A multiple-choice item that asks the test-takers to choose the most appropriate refusal responses to an invitation prompt, on the other hand, elicits offline interactional knowledge and is not tapping the intended construct, as the original activity specifies that it is verbal refusals that L2-Chinese speakers find challenging. Therefore, though both the voice message task and the multiple-choice item simplify the original activity in the TLU domain, the voice message task is more appropriate in this case since it better represents the test construct and is relatively more authentic without being too expensive to administer. This process of simplifying real-life activities to language tasks is the abstraction of the TLU domain to a Universe of Generalization (UG) (Kane, 2013). A UG represents all the possible language tasks that are modelled on the activities in the TLU domain. Backings for domain assumption 2 can be sought in the task design procedure that explicates how results of a TBNA are translated to test tasks.

Domain assumption 3 assumes that the tasks in the IC test can provide sound coverage of the TLU domain. Backings for this assumption can be gathered from inspecting the makeup of the TBNA participants, the number of tasks in the test, and the specifications of the tasks. The IC test in this book is designed to cover IC for general proficiency testing, which can include interaction at various proficiency levels in everyday, study and workplace settings. Even after the TLU domain is narrowed to a domain of more restricted interactional modes, e.g., only speaking instead of writing, this narrower domain still contains a wide range of activities, depending on test-takers' levels of proficiency. Therefore, first, the participants for the TBNA need to be carefully sampled to ensure they report crucial activities that are relevant to test-takers at different proficiency levels in the TLU domain. Second, there need to be sufficient task items in the test to represent the various activities in the domain. Third, the tasks need to vary in certain parameters such as the three contextual variables in politeness theory (Brown & Levinson, 1987) to ensure they cover different interactional situations to produce more generalizable findings. The decision of the number of tasks to include and the contextual variables to vary are informed by the TBNA conducted in this book but also by findings from previous testing research.

The last domain assumption, domain assumption 4, requires the tasks to represent activities in the TLU domain not only in terms of the authenticity of the task scenarios, but also the methods in which the tasks are delivered, and the contextual variables that shape the tasks. This assumption involves a few sub-assumptions such as, first, the task scenarios are authentic. The fact that test tasks are derived from a TBNA can boost confidence in the authenticity of the

tasks, but the authenticity of the scenarios themselves needs to be verified by objective measures. Second, the task delivery method needs to be proven to be appropriate for the task. The discussion on CMC as a test platform in Section 2.2 suggests many potential advantages of using CMC modes for task delivery. The appropriateness of matching certain CMC modes with specific IC tasks, however, needs to be empirically validated. A task that requires test-takers to criticize a close friend via voice-messaging might be flagged as unauthentic as speakers in the TLU domain might think the stakes for this social action are too high to be conducted in the voice-messaging mode. They might prefer to handle this social action via a different CMC mode such as video chat to make sure potential conflict can be immediately addressed. Therefore, validity evidence from objective measures needs to be gathered to support this assumption. Similarly, the researcher can specify the values for the contextual variables in the tasks such as close social distance or high rank of imposition, but these claims need to be grounded in objective measures apart from the researcher's judgement. The control and standardization of contextual variables are crucial to the design of test tasks as contextual variables can influence test-takers' performance. Lastly, the researcher needs to ensure the tasks can actually elicit the IC skills and behaviours that are similar to the real-world activities on which the tasks are based. In summary, domain assumption 4 is predicated on validity evidence being gathered to support these sub-assumptions to ensure the IC tasks mirror activities in the TLU domain before they are administered to test-takers.

3.1.2 The evaluation inference

Having defined the TLU domain and abstracted it to the UG which consists of IC test tasks in the domain description inference, the next inference in Figure 1, the evaluation inference, interrogates the warrant that the IC tasks can be defensibly scored to generate observed scores that measure the identified critical IC skills. This book focuses on four assumptions underpinning this warrant as listed in Table 2.

Table 2 Warrant and assumptions for the evaluation inference

Evaluation inference	
Warrant	Test-taker performance on the IC test can be scored to generate observed scores that measure critical IC skills.
Evaluation assumption 1	Task administration conditions are conducive to the elicitation of IC skills from the IC test.
Evaluation assumption 2	Sufficient training is provided to raters so that they can use the rating materials effectively.
Evaluation assumption 3	The rating scale being used by raters is able to differentiate different levels of the construct reliably.
Evaluation assumption 4	Individual raters demonstrate high intra-rater reliability.

It should be noted that at this stage of the interpretive argument, it is largely unknown what tasks are involved in the IC test. The design of the IC test will be informed by the proposed TBNA but it is reasonable to expect that the test would at least involve some interactional tasks via spoken CMC. It is inconceivable that a test that claims to measure IC in real life will not at least have some tasks that tap speakers' ability to talk. The word *talk* is used purposefully to differentiate it from the ability to *speak*, as discussed in Section 2.3. If tasks requiring talking are involved, the evaluation inference will entail issues on how the performance of talking is rated and how the rating scale functions. Operating under the assumption that talking performance will be elicited, below is a discussion on the relevant backings that are to be sought to support the four assumptions in the evaluation inference.

Evaluation assumption 1 stipulates that the test tasks are administered in a condition that can facilitate the observation of the desired test performance. This assumption is detailed in Kane et al. (1999) where the authors argued that test administration conditions should be compatible with how the test scores are to be interpreted. Backing for this assumption can be found in the test development process and test trial process. Test tasks should be trialled to make sure the test conditions are suitable for test-takers to display the critical IC skills that the test is supposed to elicit. If the IC test has a role-play task that requires the test-taker to send a voice message to their colleague over some challenging interpersonal issues, the researcher needs to ensure the test-taker only undertakes the task after they have fully comprehended what the task requires them to do. In other words, there should be no construct-irrelevant variance such as a lack of understanding of the task prompt that can interfere with the observation of

test-taker IC. Measures need to be taken to make sure that test-takers approach the tasks under the condition of maximal comprehension of the purposes of the tasks so that desired IC performance can be elicited.

Another example of considerations about the assessment condition is the preparation time given to test-takers. Asking test-takers to immediately start the abovementioned voice-messaging task right after the delivery of the test prompt is inappropriate since, in real life, test-takers would have some thinking time to contemplate what they want to say in the voice messages before sending them. Reasonable preparation time should be given to test-takers for this type of CMC tasks so that the performance elicited can reflect the IC skills this task is designed to measure.

The second evaluation assumption presupposes that raters have undergone sufficient rater training to be able to use the rating scale effectively. Rater training is crucial to the rating of productive IC test performance as IC indicators can be difficult to locate compared to traditional LC indicators. The additional challenge is that assessing IC is still a relatively novel endeavour for raters, so more training might be needed for IC rating compared to rating using more psycholinguistically-oriented LC rating scales. As discussed in Section 2.6.3, Walters (2007) highlighted the challenge of evaluating IC features even for raters who were trained in CA. In view of this, systematic rater training needs to be in place so that raters can feel confident in their evaluation of test-taker performance and their use of the IC rating scale. To facilitate rater training, the CA-informed test-taker exemplars introduced in Section 2.6 can be used to assist raters to comprehend IC indicators, locate IC indicators in test-tasker performance, and connect test-taker performance with band descriptors in the rating scale. The benefits of providing raters with test-taker exemplars and generating CA-informed exemplars have been discussed in detail in Section 2.6.3 (Davis, 2016; Greer, 2020). In summary, backings for evaluation assumption 2 can be found in rater training materials to see if potential raters have received sufficient training before they are certified as raters.

Evaluation assumption 3 requires the rating scale to be able to accurately measure test-takers' IC and distinguish different ability levels in the test-taker cohort. A fundamental goal of a language test is to precisely measure test-takers on the competence the test is designed to tap. A test is also expected to be able to differentiate a wide range of performances so that it can be used for test-takers of differing ability levels. Backings for this assumption can be first sought in how the ability range in the rating scale is arrived at. The scale needs to be developed with a wide ability range in mind. Second, the steps in the scale need to be aligned with the ability range the scale is expected to measure. The number of

steps needs to be meaningful and discernible to raters so that raters can apply the steps with confidence. These backings can be gleaned from the scale development documents to see if adequate measures are taken to ensure the scale range and scale steps are valid. In terms of raters' application of the scale, backings can be found in post-rating many Facet Rasch analysis of the rating results. Many Facet Rasch is a statistical analysis that is frequently employed in rater-mediated assessment to evaluate raters' behaviour and the functioning of the rating scale (McNamara et al., 2019). Rasch can produce a Wright map that puts test-takers and the rating scale steps along the same logit scale, which allows the researcher to inspect if the raters can use the scale to spread test-takers to different ability levels. The candidate measurement report in Rasch can indicate how precisely test-takers are being measured and how much test-taker performance fits the Rasch model's expectation. Rasch also allows for detailed analyses of scale functionality by producing a scale category functioning report that shows how the steps in the scales are being used by raters. The researcher can examine if all the steps in the scale have been sufficiently utilized to measure test-takers at different ability levels. These types of quantitative evidence of raters' use of the scale and the rating process can only be gathered and analysed once the test is administered to a large group of test-takers.

The last evaluation assumption, evaluation assumption 4, pertains to individual raters' rating reliability, which is intra-rater reliability. Intra-rater reliability is defined as how consistent a rater is in their own rating. A desirable rater demonstrates the same level of variability when they rate. Neither do they at times rate more leniently than expected nor do they rate more harshly than what a statistic model would expect. High intra-rater reliability also means raters do not display a strong halo effect or central tendency effect. A halo effect is a rater's tendency to give similar ratings to all the rating criteria for a particular test-taker's performance whereas a central tendency effect is when a rater only uses the middle steps in the scale and avoids awarding extreme scores. Both effects indicate a lack of variation in rater's rating and suggest insufficient reliability within a rater's rating. Rasch produces measures of intra-rater reliability in the form of fit values, which can be calculated after a test is administered.

3.1.3 The generalization inference

After test-takers' performance is rated to generate observed scores on a particular test in Figure 1, the interpretive argument requires evidence to support the inference that the observed scores can be generalized to universe scores on parallel tests in the UG. This inference is the generalization inference. If we go

back to the UG in the TLU domain, we can see that the items in a particular IC test are only a very limited number of items out of an infinite number of items in the UG. The test-takers undertaking one version of the IC test are also only a sample group of test-takers out of a much larger target test-taker population, which in the context of this IC test, is the entirety of L2-Chinese speakers. The generalization inference stipulates that not only the test results of a particular IC test can be generalized to other possible IC test versions consisting of parallel items, but it can also be generalized to other groups of test-takers taking the same IC test or parallel IC tests. To support this inference, four assumptions are identified in Table 3 and are to be addressed in this book.

Table 3 Warrant and assumptions for the generalization inference

Generalization inference	
Warrant	Observed scores are reflective of expected scores across parallel tasks and parallel test-taker groups in the UG.
Generalization assumption 1	There are clear test specifications documents to generate parallel tests.
Generalization assumption 2	The test-taker group is representative of the population.
Generalization assumption 3	Tasks in the IC test can reliably measure test-taker IC.
Generalization assumption 4	Different raters demonstrate high inter-rater reliability.

The first generalization assumption assumes that the researcher will create clear test specifications documents to guide the generation of similar items and the design of parallel tests. Test specifications can be viewed as an abstraction of the UG, having the potential to generate an infinite number of items within the UG. After a TBNA has identified the most challenging interactional needs that L2-Chinese speakers face in the TLU domain, the researcher can extract the specifications that cover the UG in the TLU domain so that the design of the IC test in this study follows a methodical approach. Clear and easily comprehensible test specifications also ensure that future item writers will be able to write similar items to generate parallel tests, supporting the generalization inference in the sense that it is possible to create other forms of the IC test that assess the same crucial IC skills.

Generalization assumption 2 expects that the test-taker sample for this particular IC test covers a wide range of test-takers that are representative of the target population. Since the IC test in this book is designed for L2-Chinese speakers who plan to live, study and work in China, the test-takers recruited in this study need to reflect the heterogeneity of this population. It is not ideal for example, to only sample test-takers who are undergraduate students in an Australian university as these test-takers are likely to represent a narrow age range, come from similar socioeconomic backgrounds, speak mostly English as their first language, and have similar learning needs as they are all university students. A more desirable sample would consist of test-takers from a variety of first language backgrounds, age groups and walks of life so that they have differing L2-speaker profiles and better represent the heterogenous L2-Chinese speaker population. A more diverse test-taker population also supports the argument that the test results can be generalized to a much bigger group of potential test-takers. Backings for this assumption can be sought from the test-taker recruitment procedure.

Generalization assumption 3 pertains to the assumption that the IC test tasks can offer a reliable measurement of test-taker IC. This assumption needs to be bolstered by reliability estimates, which demonstrate that the test results are reliable with different cohorts of test-takers and different parallel tests, and that the test can measure test-taker IC ability with precision. Three types of reliability estimates can be elicited to shore up this assumption. The first two are the reliability indexes produced by Rasch. Different from classical test theory, Rasch produces two reliability indexes. The first one is the Rasch test reliability, which is similar to Cronbach's alpha in classical test theory in that Rasch test reliability indicates the reproducibility of the ordering of test-takers (Winsteps, n.d., a). In other words, test reliability measures how reliable the test results are in terms of the ordering of test-takers, when the same test-takers are given a parallel IC test. Rasch test reliability underestimates this reliability while Cronbach's alpha overestimates it so the Rasch test reliability is a more stringent measure of 'person-measure-order-reproducibility' (Winsteps, n.d., a).

Rasch test reliability addresses one of the generalization requirements in that it tells us how likely the same group of test-takers will achieve similar scores if they are to take parallel IC tests made up from the infinite similar IC items in the UG. Rasch item reliability, on the other hand, provides information to address a different generalization requirement in that it informs us of how confident we can be of the item difficulties when the same IC test is given to different groups of test-takers, which is 'item-measure-order-reproducibility' (Winsteps, n.d., a). Since the target test-taker population is much bigger than the sample

test-taker cohort in a particular test-taking setting, we need to be certain that item difficulties are not going to fluctuate greatly just because a different cohort of test-takers is taking the test. This estimate is provided by Rasch in the form of Rasch item reliability. The two Rasch reliability indexes can offer us different information as to how reliable, reproducible and generalizable the test results are. They address the fundamental reliability questions in terms of whether test-takers and items with higher measures are actually stronger test-takers and more difficult items, compared to test-takers and items with lower measures.

The third reliability estimate that can be explored in the context of this book project is standard error (SE), which indicates the 'evaluations of precision' (Kane, 2013, p. 3). Since the generalization inference is interested in how observation in one test-taking setting can be generalized to other test items, other test-takers and other raters in different settings, we need to be confident that item difficulties, test-taker abilities and raters' severity indexes are measured with precision. This information can be gauged from the SEs of the related measures.

The last generalization assumption is pertinent to raters in that it assumes that different raters can award scores to test-takers consistently, which is traditionally defined as inter-rater reliability. There are different indexes of inter-rater reliability and the choice of the indexes is dependent on the rating process and the questions related to rating that the researcher wants to address (Winsteps, n.d., b). In this book, the main focus is on knowing the raters' interpretation of the steps in the rating scale, as the steps reflect different degrees of IC ability. Evidence for the generalization of raters' ratings in this particular context can be how often raters agree on their ratings and how similar raters' ratings are in terms of leniency and severity. The inter-reliability index for the former will be the exact rater agreement percentage while the one for the latter is the separation index for raters in Rasch. Both indexes can be investigated once the test is administered.

3.1.4 The explanation inference

Having arrived at universe scores from observed scores in a particular IC test in Figure 1, the next inference in the interpretive validity argument is the explanation inference, which requires the universe score to be reflective of an underlying IC test construct in the TLU domain. Introduced to the original Kane's validity argument by Chapelle et al. (2010), the explanation inference raises the question of whether test scores from IC items in the UG actually reflect an IC test construct that is modelled on the TLU domain where the UG is abstracted. Roever et al. (2014) argued that language tests frequently measure 'fuzzy, complex, intangible'

constructs, and require validity evidence that supports the connection between the test construct and the theory undergirding the construct (Roever et al., 2014, p. 146). The explanation inference requires particular attention in the validation of the IC test because in this book the test construct IC is not fully specified yet. As discussed in Section 2.4, an IC test construct can vary depending on the test format, the target second language and the TLU domain of the test. Therefore, IC testing requires a careful analysis of the connection between test performance and the test construct to ensure that the test is precisely measuring the construct it claims to measure. If indigenous criteria are used to generate an IC test construct as reviewed in Section 2.6, attention also needs to go into the connection between the test construct and theories of interaction because of the a-theoretical nature of indigenous criteria. Considering the complex nature of validating the explanation inference for this IC test, five assumptions are identified in Table 4 to support this inference.

Table 4 Warrant and assumptions for the explanation inference

Explanation inference	
Warrant	**Expected scores reflect an IC test construct in the TLU domain.**
Explanation assumption 1	The rating scale is modelled on the test construct and offers a clear representation of the test construct.
Explanation assumption 2	Test-taker test performance is reflective of the test construct.
Explanation assumption 3	The IC test construct is in keeping with theories and philosophical assumptions of interpersonal interaction.
Explanation assumption 4	The test construct measures a unidimensional construct of IC.
Explanation assumption 5	Test-taker IC performance relates to measures of LC/speaking but covers unique IC/talking variance.

The first assumption for the explanation inference is concerned with how the rating scale represents the test construct. A rating scale can be viewed as a simplified version of a test construct (McNamara et al., 2002). The relationship between the test construct and the rating scale should be transparent so that the scale in itself can explain the test construct (Knoch & Chapelle, 2017). If indigenous criteria are used in the generation of a test construct, this assumption needs backings to explain how the indigenous criteria are abstracted to a test construct, and how the test construct is mapped onto the rating scale. This

assumption also assumes the rating categories and descriptors in the rating scale are meaningful and reflective of the abilities the test construct measures. For analytic rating scales, it is common to see sub-categories within each rating category, such as the four sub-categories of fluency markers, hesitation markers, cohesive devices and topic development within the rating category *fluency and coherence* for band 9 in IELTS speaking (IELTS, n.d.). There need to be clear connections between the test construct and the sub-categories for each step in each rating category in the rating scale since the sub-categories further specify the test construct. Backings for this assumption can be found in the rating scale development documents.

Explanation assumption 2 assumes that test-taker discourse represents the test construct. If the IC test is measuring what its construct claims to measure, test-taker performance should reflect the IC skills covered in the construct. This assumption can be supported by discourse analysis of test-taker performance. When a test to be developed is measuring IC and is likely to involve interactional data, CA would be a suitable methodology for transcribing and analysing test-taker interactional data. While CA reveals the sequential aspect of talk, MCA, as discussed in Section 2.4.6, can be used in tandem with CA in the analysis of discourse to shed light on the categorial aspect of talk. Sequential-Categorial Analysis of test-taker discourse can serve as a potential backing to the assumption that the discourse-level interactional features in test-taker performance data mirror the specifics in the test construct.

The third assumption for the explanation inference requires that the IC test construct should be consistent with theoretical and philosophical understandings of interaction. As Sections 2.1 and 2.4 have explicated, different schools of thinking in philosophy, pragmatics and sociology have generated a diverse range of theories and concepts on how human interaction operates. A test construct that claims to assess the ability to handle interpersonal communication should be aligned with existing theories of interaction. Backings for this assumption are particularly pertinent in the context of test constructs informed by indigenous criteria. In Section 2.6.2 we looked at why indigenous criteria are a-theoretical as non-linguistic domain experts rarely have the theoretical understanding of interaction to inform their judgement. However, theories of interaction are abstracted from everyday practice and should be consistent with everyday-life social members' tacit knowledge of interpersonal interaction. If DEs' indigenous criteria can be shown to be in keeping with existing theories of interaction, it can validate the indigenous criteria as theory-congruent. Theory-congruent indigenous criteria can also broaden the scope of applicability of the test construct beyond the immediate context from where the indigenous criteria are

derived. This issue is particularly relevant to IC assessment as IC test constructs are usually context-specific and sensitive to the immediate local environment of the task formats and task scenarios. If the test construct in this book can be shown to be assessing higher-order IC skills that are consistent with theories of interaction, it implies that the construct to be developed is generalizable to other IC assessment tasks and contexts.

The fourth explanation assumption pertains to the dimensionality of the IC test. As reviewed in Section 2.4, there can be a wide range of potential IC indicators that are relevant to the social and interactional contexts where interaction takes place. Previous IC assessment research has located specific indicators of IC for their particular test tasks and assessment contexts, resulting in different rating categories in their analytic IC rating scales (Roever & Ikeda, 2021; Youn, 2015). Despite the various possible indicators of IC, IC as a test construct should be a unidimensional construct, which is consistent with the unidimensional assumption of assessment (McNamara et al., 2019). Therefore, the explanation inference in the validity argument is based on the assumption that the rating categories in the rating scale and the items in the test are assessing a unidimensional latent test construct. Backings for this assumption can be found in Rasch analysis of the test items and the Rasch dimensionality test on the rating categories.

The last assumption for the explanation inference, assumption 5, relates to the divergent validity of IC assessment. If an IC test claims to be measuring an IC test construct that is different from the linguistic competence (LC) construct, there needs to be evidence to support the differentiation between IC and LC in test-taker performance. This brings us back to the theoretical and operational discussions on talking/IC versus speaking/LC in Sections 2.3 and 2.4. A test that is designed to measure talking/IC needs to be able to demonstrate that it is not assessing speaking/LC. There is of course always going to be an overlap between IC and LC as LC represents the linguistic devices that speakers can mobilize to conduct interaction. However, for an IC test to be a valid assessment instrument, it needs to show that it assesses variance not already accounted for by LC tests. From a more theoretical perspective, the connection between IC and LC is also pertinent to how the competence of L1 speakers is defined in the test construct. If a test is measuring LC, L1 speakers by default should be the benchmark on which the rating scale is based as L1 speakers possess the ideal LC in traditional psycholinguistic speaking frameworks. If a test is measuring IC, we can no longer safely assume that L1 speakers are necessarily always the most appropriate, effective and successful communicators. Conventional wisdom informs us that there can be L1 speakers who are not apt at using their near-perfect linguistic

skills to conduct effective interaction. If this assumption regarding L1 speakers' IC is built into the explanation inference, there needs to be backings supporting the assumption that L1 speakers are not invariably the best in IC. Taking into account the complexity in supporting explanation assumption 5, both qualitative and quantitative backings need to be sought. For qualitative backings, discoursal data from test-taker performance can be inspected to better understand the connection between IC and LC. For quantitative backings, the relationship between test-takers' scores on the IC test and their scores on LC measures can be examined. L1 speakers should also be included in these analyses to see if an IC test is indeed measuring IC or is just another measure of LC.

3.1.5 The extrapolation inference

If the observed scores from the IC test can be generalized to the universe scores in the UG and if the universe scores can be attributed to an underlying IC test construct in Figure 1, the next inference in the interpretive argument is the extrapolation inference. This inference states that the universe scores in the UG can be extrapolated to measures of activities in the TLU domain where the UG is abstracted. This inference is crucial to the inference chain in the interpretive argument as language tests are supposed to be modelled on language abilities that are observed in real-world activities. As Section 2.3 illustrates, if test results of a language test cannot be extrapolated to real-world behaviour, the resultant test scores are not valid indicators of test-takers' communication abilities *in the wild*, which will threaten the usability and validity of the test as a whole (O'Sullivan, 2019). Three assumptions are identified in Table 5 that support this inference in the IC test to be developed in this book.

Table 5 Warrant and assumptions for the extrapolation inference

Extrapolation inference	
Warrant	**The IC test construct reflects the quality of test-taker interaction in the TLU domain.**
Extrapolation assumption 1	The test construct embodies the evaluation of critical IC skills in the TLU domain.
Extrapolation assumption 2	The IC rating categories measure test-taker IC in activities in the TLU domain that resemble the assessment tasks in the UG.
Extrapolation assumption 3	Stakeholders perceive the IC test as measuring test-takers' ability to interact in the TLU domain.

The first assumption for the extrapolation inference assumes the universe scores in the UG represent the domain scores in the TLU domain. The implication here is that judgement on test-taker IC test performance reflects how their performance is assessed in the TLU domain, which, for this book is everyday life, work and study settings. Backings for this assumption can be found in DEs' judgement in the development of the IC test. As Section 2.6.2 reveals, using DEs in the creation of rating scales follows the rater-perception tradition of scale development. However, DEs represent a special group of raters in that they are usually not linguistically trained but have rich experience in and knowledge of the TLU domain (Knoch, 2012; Pill, 2016; Sato & McNamara, 2019). Knoch and Chapelle (2017) argued that using DEs' judgement in scale development can assist test designers to develop scales that reflect how test-taker language use is assessed in the TLU domain, strengthening the extrapolation inference. Different from judgement criteria adopted by linguists, testing specialists or language teachers, DE criteria can more accurately represent how test-taker talk is informally assessed by everyday-life members of society in everyday-life settings. Sato and McNamara (2019) also contended that relying solely on linguists' judgement of effective communication can be restrictive because, in real life, L2 speakers are more frequently judged by everyday-life members such as their friends, classmates, colleagues and employers, most of whom are non-linguists. The extrapolation inference can be supported if everyday-life DEs' judgements are elicited to inform the development of the test construct and the design of the rating scale.

Using DEs' indigenous criteria has two additional advantages in relation to the extrapolation inference. As reviewed in Section 2.6.1, IC research has traditionally relied on IC researchers' judgement on what IC indicators to use to index L2 speakers' IC. Such an approach has led to a proliferation of potential IC markers that are either applicable across many different assessment contexts (e.g., turn-taking), or relevant to some very specific task formats and contexts (e.g., preliminaries in launching a refusal). The vast number of potential IC markers was well captured in the IC tree in Galaczi and Taylor (2018). The indigenous criteria from DEs can inform researchers as to what IC indicators to prioritize since DEs' judgement reflects what everyday-life interactants consider important to interactional success. This ensures that the test is actually measuring what interactants in the TLU domain value, strengthening the extrapolation inference.

The other advantage of recruiting DEs is that DEs have insight into the real-world consequences and implications of test-takers' interaction. Possessing knowledge of how interaction should unfold and the consequences of appropriate or inappropriate interaction, DEs can guide the researcher in establishing the connection between test performance and real-world consequences. For example,

DEs may inform the researcher that a particular type of test-taker performance in real-life will be highly problematic and engender serious repercussions for test-takers. More training and teaching resources can be subsequently developed to address such issues, strengthening the link between assessment and pedagogy. This argument echoes the one made in Sato and McNamara (2019) when the authors maintained that linguistic laypersons are most of the time the ultimate judges of language performance in the TLU domain. Eliciting DEs' judgement can therefore inform the researcher and end-users of the IC test in terms of the real-world implications of test-taker performance when it is extrapolated to non-testing activities in the TLU domain.

While extrapolation assumption 1 assumes that the test construct reflects how test-takers are being judged in the TLU domain, extrapolation assumption 2 presupposes the test is measuring test-taker performance in similar activities in the TLU domain. If the universe scores in the UG are to be extrapolated to domain scores in the TLU domain in Figure 1, the researcher needs to be confident that the measurement of IC abilities in the test tasks is comparable to the one of IC abilities in similar activities in real life. Chapelle et al. (2008) used test-taker self-assessment and instructor rating as backings for this assumption. A similar approach can be adopted in this book by exploring how much test-taker test performance correlates with their self-assessment of their interaction in everyday life. Peer assessment can also be elicited to see if test-taker test scores reflect how people that frequently interact with test-takers perceive of test-takers' IC in the TLU domain.

The last assumption for the extrapolation inference pertains to how stakeholders perceive the extrapolative strength of the IC test. End-users of the test such as language teachers and company employers are all involved in the test use in the TLU domain and can assess how accurately the test scores reflect a speaker's ability to interact in the TLU domain. One group of stakeholders is test-takers themselves. As participants in the test and users of test results, test-takers have first-hand experience of their cognitive processes when taking the test. They can assess how likely their real-world behaviour in similar activities in the TLU domain is consistent with their behaviour in the test settings. Therefore, one form of backing for extrapolation assumption 3 is to see whether test-takers think their test performances actually reflect their behaviour in the TLU domain.

3.2 The design of the three studies

Having identified the assumptions for each of the inferences that are pertinent to the particular context of the IC test for this book project, the design and

validation of the IC test in this book is structured around seeking backings for the assumptions. Considering the different types of evidence needed to support the assumptions, the test design and validation are staged in three studies. Each study is designed to address a number of the assumptions in the interpretive argument. The research questions for each study are the critical evaluations of the backings for each assumption. In other words, the research questions can be phrased as: are the backings sufficient to support this assumption?

It should be noted that the studies do not follow the inferences in a linear fashion. A study can address assumptions from different inferences but taken together, the three studies seek to gather backings that support all the assumptions identified in Section 3.1, generating a coherent validity argument. The three studies are discussed separately in Chapter 4, Chapter 5 and Chapter 6 with their own methodologies, results and initial discussions. Chapter 7 brings the results of the three studies together in the framework of a validity argument to evaluate the validity of the IC test. The following three sections, Section 3.2.1 to Section 3.2.3, sketch the purposes each study serves and the assumptions each study addresses.

3.2.1 Study one, relevant assumptions and research questions

Study one is concerned with the test development phase of the book. The assumptions that study one aims to address are listed below in Table 6. The description of the set-up of study one follows the order of the assumptions.

Table 6 Research questions for study one

Research questions in study one based on assumptions	
Assumptions	**Assumptions reformulated as questions**
Domain assumption 1	Can critical activities indexing IC be identified in the TLU domain?
Generalization assumption 1	Are there clear test specifications documents to generate parallel tests?
Domain assumption 2	Can IC assessment tasks be developed that are based on the identified critical IC activities?
Domain assumption 3	Do the IC tasks in the test offer sound coverage of the TLU domain?
Evaluation assumption 1	Are the task administration conditions conducive to the elicitation of IC skills from the IC test?
Domain assumption 4	Do the task scenarios and delivery methods represent activities in the TLU domain?

In view of the assumptions and research questions, study one starts with a task-based needs analysis (TBNA) on the most challenging interactional situations that L2-Chinese speakers frequently struggle with. This will narrow down the TLU domain and identify the most crucial IC skills and most demanding interactional situations in the context of L2 Chinese, generating backings for domain assumption 1. Based on the identified IC skills and IC scenarios, I can define the specifications of the IC test. The quality of the specifications documents can be assessed to see if they sufficiently support generalization assumption 1. Based on the results of the TBNA and test specifications, I can develop prototypical IC test items. The process of item construction can be evaluated to see if there is sound backing for domain assumption 2. Once a sufficient number of items are developed based on insight from well-sampled TBNA informants and in accordance with the test specifications, I can assess if the TLU domain is adequately represented by the items in the test, which addresses domain assumption 3. The next step is to pre-pilot and trial the items to see how the items should be delivered to achieve optimal elicitation of the target IC skills. Other language testing specialists can assess if the task administration conditions I have conceived can provide enough backing for evaluation assumption 1. Finally, the items in the test need to be validated in terms of the appropriateness and authenticity of their content, delivery method and contextual variables. Both qualitative and quantitative backings can be sought to support this assumption, which is domain assumption 4.

The writing and reporting of study one in Chapter 4 largely follow the structure mentioned above. A concern that is related to study one but does not correspond to any inferences or assumptions in the interpretive argument is the use of computer-mediated communication (CMC) in testing. As Section 2.2 explains, CMC can improve the ease of test delivery and the practicality of testing, especially for highly interactional tests like IC tests where productive skills are assessed and when F2F testing is too expensive or impractical. The use of CMC in this book will be discussed and investigated in study one, though it does not fit into any of the abovementioned assumptions or the argument-based validity framework in general.

3.2.2 Study two, relevant assumptions and research questions

Study two is mainly concerned with the design of rating materials for the IC test developed in study one. As a rating scale is a simplified representation of the test construct, study two also addresses inferences beyond the evaluation inference. The assumptions and research questions that study two concerns are listed in

Table 7. Similar to study one, the description of the stages in study two follows the ordering of the assumptions.

Table 7 Research questions for study two

Research questions in study two based on assumptions	
Assumptions	**Assumptions reformulated as research questions**
Explanation assumption 4	Is the IC test construct unidimensional based on pilot testing?
Extrapolation assumption 1	Does the test construct embody the evaluation of critical IC skills in the TLU domain?
Explanation assumption 1	Is the rating scale modelled on the test construct and offering a clear representation of the test construct?
Evaluation assumption 3	Can the steps in the rating scale precisely measure test-takers' different ability levels?
Explanation assumption 2	Is test-taker performance reflective of the test construct?
Explanation assumption 5	Does test-taker IC performance relate to measures of LC/speaking but cover unique IC/talking variance?
Evaluation assumption 2	Are raters well trained to use the rating materials?
Explanation assumption 3	Is the IC test construct in keeping with theories and philosophical assumptions of interpersonal interaction?

After the pre-pilots and trials of the IC test in study one, the IC test should be in a shape that is suitable for pilot testing. At the start of study two, I first conduct a pilot test of the IC test on a small group of test-takers. The pilot test serves two purposes. First, it generates preliminary psychometric measures of the test to assess how the test, pilot test-takers, items and raters function. This information can assist me with evaluating if the test is indeed assessing a unidimensional test construct before the test construct is specified. If the preliminary analysis of pilot test results suggests the test is measuring a unidimensional construct, this can serve as a piece of evidence supporting explanation assumption 4, though more evidence for unidimensionality will be sought in the main testing in study three.

The second purpose for the pilot test is that, if the latent test construct is found to be most likely unidimensional, pilot test-taker sample performances can serve as stimuli for everyday-life DEs to generate their indigenous IC criteria, which addresses extrapolation assumption 1. The indigenous criteria elicited can be analysed to produce the blueprint of an IC test construct. The procedure involved

in the generation of the indigenous criteria, the conversion from the indigenous criteria to rating categories in the IC rating scale, and the development of the rating scale can be assessed to see if there is sufficient backing for explanation assumption 1. The main backing for evaluation assumption 3 will be sought during the main testing stage in study three, as it requires the main testing dataset to be analysed quantitatively. However, one important piece of evidence for evaluation assumption 3 that can be gathered in study two is how the steps in the rating scale are developed. The steps in the rating scale have a direct impact on the range of abilities measured in the IC test. Clearly defined steps can assist raters with differentiating performances at different levels, improving the precision of measurement.

The test-taker performance elicited from the pilot test in study two can also be used to seek backings for the remaining assumptions pertinent to study two. Test-taker discourse can be transcribed and analysed via Sequential-Categorial Analysis (combining CA and MCA) to examine if interaction at the micro-level reflects the test construct, which is the evidence needed for explanation assumption 2. The fine-grained analysis of test-taker performance can also generate qualitative evidence, which sheds light on whether the IC construct covers variance beyond traditional LC frameworks, just like what Lee and Hellermann (2014) and Jenks and Brandt (2013) have shown in Section 2.4.2. This type of qualitative evidence can be used as backing for explanation assumption 5.

In addition, the refined Sequential-Categorial Analysis (using both CA and MCA) of test-taker performance excerpts can potentially be transformed into CA-informed exemplars (Greer, 2020) for rater training if the performance excerpts are carefully selected to represent the test construct. Greer (2020) has shown the applicability of CA-informed exemplars in rater training. Carefully designed rater training materials are a prerequisite for effective rater training, the provision of which supports evaluation assumption 2.

Finally, the a-theoretical nature of indigenous criteria implies that more work is needed to shore up the theoretical underpinning of an IC test construct derived from indigenous criteria, compared to test constructs that are directly mapped from existing theories and frameworks of speaking/LC. The last step in study two will be to position DEs' indigenous criteria under existing theories of interpersonal interaction to ascertain the theoretical soundness of the IC test construct. This process has the potential to expand the test construct beyond its immediate test context to cover a wider range of assessment formats and situations. This is built on the assumption that DEs' criteria are reflective of more holistic concepts and theories of interaction, which is explanation assumption 3.

3.2.3 Study three, relevant assumptions and research questions

Study three is the main testing study when the IC test and related instruments are administered to the main cohort of test-takers. This study addresses the assumptions listed in Table 8.

Table 8 Research questions for study three

Research questions in study three based on assumptions	
Assumptions	**Assumptions reformulated as research questions**
Generalization assumption 2	Can the test-takers be sampled to approximate a sound representation of the real-world test-taker population?
Evaluation assumption 2	Are raters well trained to use the rating materials?
Evaluation assumption 3	Can the steps in the rating scale precisely measure test-takers' different ability levels?
Generalization assumption 3	Can tasks in the IC test reliably measure test-taker IC?
Evaluation assumption 4	Do individual raters demonstrate high intra-rater reliability?
Generalization assumption 4	Do different raters demonstrate high inter-rater reliability?
Explanation assumption 4	Is the IC test construct unidimensional?
Explanation assumption 5	Does test-taker IC performance relate to measures of LC/speaking but cover unique IC/talking variance?
Extrapolation assumption 2	Do the IC rating categories measure test-taker IC in similar real-life activities in the TLU domain?
Extrapolation assumption 3	Do stakeholders perceive the IC test as measuring test-takers' ability to interact in the TLU domain?

Backings for the first two assumptions, generalization assumption 2 and evaluation assumption 2, can be sought in the methodology of study three. Generalization assumption 2 requires the careful selection of test-takers to ensure their profiles are diverse enough to allow for the generalization of one test-taker group to the entire test-taker population. Evaluation assumption 2 requires that sufficient training is provided to raters of the main testing study so that they can use rating materials confidently.

Once the IC test is administered and test-taker scores are generated after rater rating, potential backings for the following five assumptions, from evaluation

assumption 3 to explanation assumption 4, can be sought from Rasch analyses of test-taker scores. Rasch generates indexes that can be used to assess raters' use of the rating scales (evaluation assumptions 3), the reliability of test tasks (generalization assumption 3), the reliability of rater rating (evaluation assumption 4 and generalization assumption 4) and the dimensionality of the latent IC test construct (explanation assumption 4).

Backings for the last three assumptions, from explanation assumption 5 to extrapolation assumption 3, require information that cannot be generated by the IC test alone. Study two has the potential to elicit qualitative discoursal evidence to support explanation assumption 5, which is explained in Section 3.2.2. However, if quantitative evidence is to be generated for this assumption, such as the evidence in Ockey et al. (2015) in Section 2.4.2, measures of LC need to be in place so that test-takers are assessed both on LC and IC. This will make the evaluation of the relationship between their LC and IC possible from a quantitative perspective. To examine backings for extrapolation assumption 2, self and peer-assessment questionnaires need to be developed and administered to test-takers and test-taker peers to see if test-taker performance on the test mirrors their performance in similar activities in real life. Backing for extrapolation assumption 3 requires the elicitation from test-takers of their perception of the extrapolative strength of the IC test to their real-world conduct. Questionnaire items can be developed to measure test-taker perception.

In addition to the assumptions identified as relevant to study three, study three also provides the opportunity to elicit test-takers' attitudes to the IC test. As stakeholders and end-users of the IC test, test-takers' opinions towards the IC test impact on the take-up and marketability of the test. Therefore, when the IC test and other questionnaire items are administered to the test-takers, some questionnaire items that assess the construct of test-taker attitude towards the test are included. Though stakeholder uptake is not included in Kane's validation framework, it is an important piece of information to gather for test design as it pertains to the economic prospect of the test, which is a concern for large testing organizations.

Chapter 4 Study one: Task-based needs analysis and test design

Chapter 4 reports on the first study conducted for this book, which focuses on the design of the IC test. The rationale for conducting study one and addressing the research questions of study one is detailed in Section 3.2.1.

In order to develop an IC test, I first conducted a task-based needs analysis (TBNA) to identify what L2-Chinese speakers struggle with the most when it comes to interpersonal interaction, eliciting insight from L2-Chinese speakers, L1-Chinese teachers and L1-Chinese interactants. Thematic analysis revealed that L2-Chinese speakers found the management of disaffiliative social actions most challenging, which were used as the basis for creating test tasks. Concomitant interactional issues such as sociopragmatic concerns, pragmalinguistic concerns, interactional features, content knowledge and linguistic issues were also incorporated into the test tasks, offering comprehensive coverage of the challenges identified in the TBNA. Drawing on findings from the TBNA and the rich critical incidents reported by TBNA participants, I drew up the test specifications and generated draft items. The draft items were then submitted to an iterative process for validation, which involved feedback from L2-Chinese informants, L1-Chinese informants, applied linguist informants and testing specialist informants. The four groups of informants participated in interviews, panel discussions, trials, and pre-pilots to examine the quality of the items. I also assembled an item review committee and conducted two norming sessions to further gather validity evidence for the items. The resultant IC test is a nine-item test that has three interactive modes, three Power variables from politeness theory, and three sub-target language use (TLU) domains. The test is designed for the computer-mediated communication (CMC) environment, which increases test practicability. The test specifications are presented in Section 4.2.2. Section 4.2.5 offers an overview of the test and the task shell. The needs analysis part of study one was reanalysed and published in Dai (2023a) where I present a new IC needs model to conceptualize L2 speakers' interactional demands. More qualitative analysis of needs analysis informants' insight is also provided. Interested readers can refer to Dai (2023a) for details.

4.1 Methodology of study one

Section 4.1 explains the methodology involved in the TBNA and the test development process that flows out of the TBNA. The TBNA incorporated a range of triangulation methods as suggested by Brown (2009) and explained in Section 2.5.3. Both triangulations of information sources and data elicitation methods were adopted.

After the TBNA was conducted, I started the design of the test by first developing draft items based on findings from the TBNA and the critical incidents reported by TBNA participants. Then I involved four groups of informants in the review and revision of the draft items. A more polished set of items was submitted to an item editing/moderating committee consisting of language assessment experts for feedback. To further validate the items, two norming sessions were conducted with L1-Chinese speakers to examine the functioning and quality of the items.

Since study one consists of two stages – the TBNA and the subsequent test design process – I separately report participants, instruments, procedures, and data analysis techniques for the TBNA and the test development process from Section 4.1.1 to 4.1.4.

4.1.1 Participants

4.1.1.1 TBNA participants

I recruited 18 participants for the TBNA in this study, as Figure 2 illustrates. Participants fell under three categories: L2-Chinese speakers of different degrees of socialization in the target community, Chinese language educators from three different course levels, and L1-Chinese speakers with frequent interaction with L2-Chinese in three sub-TLU domains: study, work, and everyday interaction. Tables 9–11 provide details of the 18 participants in the three groups, which show that the participants were selected to represent different angles, perspectives, and contexts so as to offer comprehensive coverage of the target domain. Although the number of participants was not particularly large, the consideration that went into participant selection ensured the selected participants' reporting encompassed activities in the language use domain.

Figure 2 TBNA participant information

Note. F for female, M for male, pseudonyms used, asterisks indicating the participants had applied linguistics training.

Participant selection in this study embodied a few triangulation strategies in terms of sources of information (Brown, 2009). First, there was a range of observing angles represented. Through consulting the perspectives of L2 speakers, L1 speakers and language teachers, I gained a better understanding of where L2 speakers saw themselves struggle most, where teachers saw their students struggle most and where L1 speakers saw L2 speakers struggle most. The perspective of L1 speakers was particularly valuable and underrepresented in previous TBNA studies. L1 speaker interactants do not have a pedagogical intent as teachers do. On the contrary, they observe L2 speaker language use *in situ* and form intuitive insight, which can enrich the results of a TBNA.

Table 9 Details of the TBNA L2-Chinese speaker group

High proficiency	Alexa*	L1 Italian. Age not disclosed. More than 10 years' experience studying Chinese as a foreign language outside China. Completion of a master's degree in Chinese Philosophy and Translation in a university in China with Chinese as the medium of instruction. More than 10 years' experience living in China. More than 3 years' experience working in China. Attainment of HSK 6.
	Dave*	L1 British English. Age range 26–30. 3–5 years' experience studying Chinese as a foreign language outside China. Completion of a bachelor's degree in East Asian studies in China with Chinese as the medium of instruction. More than 10 years' experience living in China. More than 3 years' experience working in China. Attainment of HSK 6 and HSKK[2] advanced.
Medium proficiency	Crissy	L1 Australian English. Age range 26–30. 5–10 years' experience studying Chinese as a foreign language outside China. 5–10 years' experience living in China. 1–3 years' experience working in China. Attainment of HSK 5.
	Mike*	L1 Australian English. Age range above 40. Naturalistic learner. More than 10 years' experience living in China. PhD candidate working on interpersonal communication in Chinese. Married to an L1-Chinese partner. More than 3 years' experience working in China.
Low proficiency	Ella and Dan	L1 Australian English. Age range 21–25 for Ella, 26–30 for Dan. 3–5 years' study experience of L2 Chinese outside China. One year's experience living in China. Below one year's internship and work experience in China. Lower HSK levels achieved but not disclosed.

2 HSKK: Hanyu Shuiping Kouyu Kaoshi (Spoken Chinese Proficiency Test), a speaking test that is separate from HSK and that has three levels: beginner, intermediate, and advanced.

Table 10 Details of the TBNA L1-Chinese educator group

Advanced	Wang Ni and Deng Yang*	Teaching and coordinating advanced-level Chinese language degree programs in tertiary settings. Frequent interaction with high proficiency L2-Chinese students in class and outside class. Deng Yang has a PhD in Applied Linguistics while Wang Ni has a PhD in Chinese Literature.
Intermediate	Chu Song* and Zhou Wu	Teaching and coordinating Chinese language non-degree courses. Experience in teaching and tutoring intermediate-level L2-Chinese students. Experience in interaction with students outside classroom settings (e.g., socials and course-related travels). Chu Song has master's degrees in Applied Linguistics and Teaching Chinese as a Foreign Language.
Beginning	Hu Yin and Ma Dong	Teaching and language program coordination experience with L2-Chinese students of beginner-level proficiency. In-depth knowledge of beginner-level students' interactional needs from one-on-one tutoring.

Table 11 Details of the TBNA L1-Chinese interactant group

Work	Dai Ling*	Professional Chinese-English interpreter working in a multilingual translation service. Experience interacting in Chinese with L2-Chinese colleagues and business partners from a range of L1 and professional backgrounds.
	Zhu Huang	Employed in an internship unit that connects L2-Chinese professionals to internship opportunities in China. On-the-ground experience liaising with L2-Chinese working professionals on a range of work-related issues in China.
Study	Xuan Zhong	Science major undergraduate student in a university in China. Frequent experience interacting with L2-Chinese students who come to their university through exchange programs. Provision of assistance to L2-Chinese students with study in China.
	Li Hu*	Extensive experience interacting with L2-Chinese speakers in tertiary educational settings. PhD candidate focusing on interaction in Chinese.
Life	Ke Han	Language enthusiast interested in making friends with L2-Chinese speakers and learning about their languages and cultures. Frequent social interaction with L2-Chinese speakers in a range of everyday informal settings.
	Song Mu	Life partner of Mike, who was a participant in the L2-Chinese speaker group. Daily experience in interacting with Mike using Chinese. Frequent observation of how Mike interacts in Chinese with her extended L1-Chinese-speaking family and L1-Chinese friends.

Second, there were different degrees of target community socialization within the L2 speaker group, different course levels within the L1 teacher group, and different sub-TLU domains within the L1 speaker group. This assisted me to obtain not only in-depth knowledge of the differing needs of L2 speakers at different stages of language acquisition but also a more comprehensive picture of language use in the three most representative domains. Although the participants for the L2-Chinese speakers were mainly from English-speaking backgrounds, this bias was mitigated by insight provided by participants from the other two groups, who had frequent interaction with L2-Chinese speakers from a wide range of L1 backgrounds.

Third, there was a balanced representation of gender in every sub-participant group. Linguistic and sociolinguistic research on Mandarin and Cantonese speakers has noted gendered linguistic practices such as the use of sentence-final particles in Chinese communities (Chan, 1996, 1998; Diao, 2014; Farris, 1988, 1994). The similar number of male and female participants in this study mitigated the effect of gender bias in participants' reporting.

Finally, there was a mix of differing domain expertise. Compared to participants whose names are not marked by an asterisk in Figure 2, those with asterisks received formal post-graduate training in areas related to interpersonal communication such as cultural studies, translation and interpreting, applied linguistics and CA. Similar to Serafini et al. (2015), this study made a conscious effort by eliciting information from both general informants and trained applied linguists, whose domain expertise offered illumination on interactional issues sometimes too subtle to be noticed by untrained eyes.

4.1.1.2 Test design participants

After the TBNA was completed, I wrote up draft items and planned the test design process which included two stages: item revision and item norming. In terms of participants for the two stages, I first recruited participants as informants to provide feedback on item editing and moderation. After the draft items were modified and semi-finalized, I recruited L1-Chinese participants for two norming sessions to validate the specifications of the items. Participants for the two stages of test design are reported below separately.

Item review and moderation participants

After I developed draft test items based on findings from the TBNA, I recruited four groups of informants to inform item review and revision. The four groups were L1-Chinese speakers, L2-Chinese speakers, Chinese language teachers

and applied linguists/testing specialists. These four groups of participants are separate from the TBNA informants described in Section 4.1.1.1.

The L1-Chinese speaker informants were all residents of China and were able to comment on whether the item scenarios were authentic or realistic based on their lived experiences using Chinese in China. They were also able to comment on the suitability of the matching between the task requirement and the task method prescribed, such as whether it was feasible or common to use voice messaging as a task delivery method for the launching of certain social actions.

The L2-Chinese speaker informants provided feedback on whether the language in the task scenarios was accessible to L2 test-takers at different proficiency levels. During interviews with participants from this group, they were able to provide alternatives to the grammatical structures, lexical choices and interactant names used in the scenarios to make the test tasks comprehensible to L2 test-takers with lower intermediate proficiency in L2 Chinese.

The Chinese language teachers provided similar feedback as the L2 speakers did, but their insight was more from a pedagogical perspective. Apart from assisting with revising the prompts in the test tasks, the Chinese teacher group offered comments on the selection of task scenarios to make the tasks more aligned with the learning needs they had identified in students.

Finally, the applied linguists/testing specialists group reviewed the items at various stages of item development from a research and test operationalization perspective. In this process I purposefully invited applied linguists with backgrounds in pragmatics, CA and IC to assess if the items were well designed to elicit the intended social actions and interactional features. The testing specialist group provided feedback on test delivery, the internal structure of the test, potential statistical analyses to be applied and other testing-related concerns.

Participants from the four groups also trialled the items at various stages of item design so as to offer more targeted feedback. This is in line with the proposition put forward in Alderson et al. (1995) where the authors contended it is desirable for informants to trial and answer the draft items so as to better assess the functioning of the items.

Norming session participants

After the test items were reviewed and trialled by the four groups of informants, the items to be included in the test went through two norming sessions to ensure they functioned as intended with larger numbers of participants. 33 L1-Chinese speakers participated in the first round of norming. The selection criteria for the 33 participants were that they had to be L1 speakers of Chinese,

grew up in China, completed tertiary education in China and were working in China at the time of participating in the research. To mitigate the influence of different interactional norms, I required that the participants needed to have been continuously residing in China without any residence experience outside China apart from short-period travels. These requirements ensured the norming participants were a relatively homogenous group that could reliably assess the contextual pragmatics variables in the test tasks. After the first round of norming was completed, I revised the items and administered them to a second group of norming participants. 21 L1-Chinese speakers who met the same selection criteria of the first norming session participated.

4.1.2 Instruments

4.1.2.1 TBNA instruments

In the TBNA phase of study one, two data elicitation methods were used: hermeneutic-Socratic (H-S) interviews and longitudinal reflective diaries. The two methods were adopted to triangulate data elicitation methods, on top of the various triangulation strategies employed for the information sources as detailed in Section 4.1.1.1.

Hermeneutic-Socratic interviews

Though interviews are frequently used by needs analysts to elicit knowledge of learner needs (Chaudron et al., 2005; Cowling, 2007; Huh, 2006; Lambert, 2010; Malicka et al., 2017; Serafini, et al., 2015), little existing TBNA literature has moved outside traditional scripted interviews and semi-structured interviews. Interview methodologists, however, have long questioned the epistemological underpinning of such practices, where interviews are reduced to a process of asking the right questions and expecting willing interviewees to proffer enlightening answers (Roulston, 2010). H-S interviews break away from this tradition by positioning both the interviewer and interviewees as active co-participants in the quest for meaning (Dinkins, 2005). Drawing on the Heideggerian sense of phenomena and the Socratic approach to dialogue, H-S interviewing focuses on defining commonplace but highly abstract concepts (e.g., what is successful interaction to you), contextualizing questions with analogies (e.g., there was this time when I did…) and orienting to conflicts in the dialogue (e.g., before you mentioned A, but now you changed to B). Such techniques are particularly helpful in unearthing information on complex topics such as challenges in interaction and what makes interaction challenging. The H-S interview protocol used in study one can be found in Appendix I. From reading

the protocol readers can notice how H-S interviewing suspends pre-conceived notions on the side of the interviewer and avoids asking leading questions, such as 'do you find apologizing in Chinese difficult?', which are not uncommon in existing TBNA studies. Another advantage of using H-S interviews is that the rich contextualized information elicited from participants' reporting of critical incidents can assist researchers with developing authentic language tasks (Long, 2013b). This point will be illustrated with an example in Section 4.2.3.

Longitudinal reflective diaries

Though the critical incidents generated from H-S interviews have depth in meaning, they lack breadth, which is a renewed and renewing understanding of the same phenomenon. A longitudinal reflective diary, as a secondary and complementary information elicitation method, addresses this deficit in the quality of data from H-S interviews. H-S interviews marked the start of an iterative process where both the interviewer and interviewees began to ponder over what L2-Chinese speakers found most challenging in their interaction. After the initial H-S interviews, I encouraged participants to keep reflecting on the interactional challenges discussed in the H-S interviews and maintain a diary to note down their reflections and any additional critical incidents they noticed. This period lasted three months, which generated additional data that is longitudinal in nature. In a sense, the initial H-S interviews heightened participants' awareness of the interactional plights faced by L2-Chinese speakers, while the longitudinal reflective diary encouraged them to continue to notice and contemplate on this topic. Given that all participants were in-service participants, which means they were working, studying and living in Chinese-speaking communities at the time of participating in study one, they had frequent opportunities to observe and engage in interaction in Chinese, reflect on their practice and experience, and generate rich information for the TBNA. Additional critical incidents noticed in this period in their reflective diaries were reported to the researcher.

4.1.2.2 Test design instruments

Norming questionnaires

After the items were developed from the TBNA and revised based on feedback from the four sources of informants in the test development process, I conducted two rounds of norming sessions using norming questionnaires. The purpose of the norming procedures was to ascertain: (1) if L1-Chinese speakers would use the intended social actions in the test tasks as I expected, (2) if they found the test tasks authentic and realistic, (3) if they deemed the task delivery methods

appropriate for the respective tasks, and (4) if they agreed on the variable settings (power, distance, and imposition) in the tasks as preconceived by the researcher. The norming questionnaire contained six questions. The first question was similar to a discourse completion task (DCT) that asked the questionnaire respondents to write down what they would do or say in the given scenario. Although research has shown that DCTs cannot elicit the same speech performance as in more naturalistic tasks such as role plays (Golato, 2003), DCTs were considered appropriate for this step in study one as the norming procedure was only to roughly assess if respondents oriented to the test tasks as I expected, instead of analysing the exact language questionnaire respondents would use in the task scenarios. The questionnaire in both the original Chinese version and the English-translated version is included in Appendix II.

4.1.3 Procedures

4.1.3.1 TBNA procedure

Figure 3 illustrates the procedure of the TBNA. I first sampled representative participants based on the four pre-determined criteria for information source triangulation as described in Section 4.1.1.1. After the first H-S interviews, I maintained written communication with participants and collected insight from participants' reflective diaries over the following three months. Follow-up H-S interviews using the same interview protocol were set up on an *ad hoc* basis to clarify confusing critical incidents reported in participants' diaries. Interviews were audio-recorded and transcribed while written data were extracted from participants' diaries. Two L1-Chinese speakers were recruited to listen to randomly selected excerpts of the data in order to verify correct transcription and appropriate data management. Making source data available for inspection helped to strengthen the basis for data analysis and constituted a form of audit trail (Brown, 2001; Seale, 1999). Participants also received a copy of their interview transcript for member checking, a process that further validated my understanding of source data (Roulston, 2010). In situations where participants voiced disagreement with the information in their interview transcripts, I provided raw data for inspection. Further discussions usually followed, leading to more refined interpretations of the reported critical incidents. Both the audit trail and member checking increased the validity and verifiability of data processing and were used again at the data analysis stage of this study.

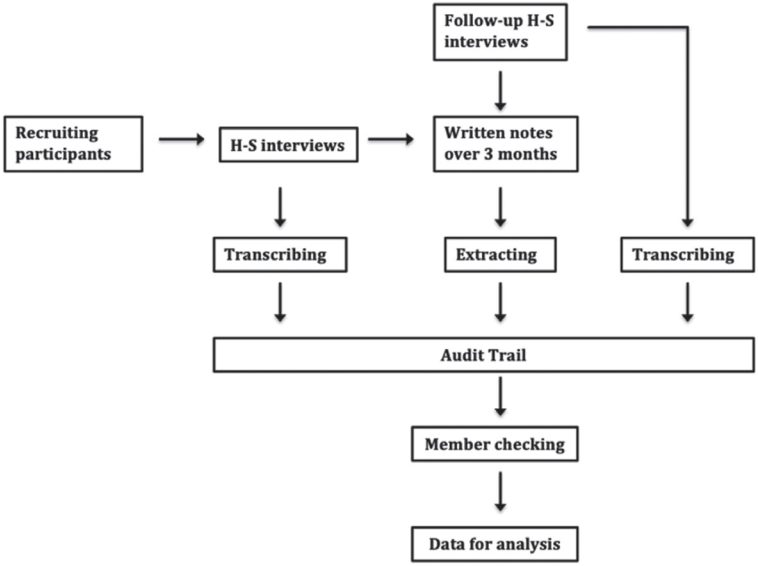


Figure 3 Procedure of the TBNA

4.1.3.2 Test design procedure

Having conducted the TBNA to identify where L2-Chinese speakers struggled the most in terms of L2-Chinese interaction, I developed the test specifications (see Section 4.2.2) from the TBNA results. Draft items were written by me based on the critical incidents reported by the participants in the TBNA.

The revision of the draft items and selection of draft items for the IC test were an interactive process, involving multiple stages of qualitative and quantitative validation of the items. The qualitative validation was sought from interviews, panel discussions, trials, and pre-pilots with participants in Section 4.1.1.2 who were able to provide insight into the suitability of the draft items. The four groups of informants in Section 4.1.1.2 – L2-Chinese speakers, L1-Chinese speakers, L1-Chinese teachers, applied linguists and testing specialists – contributed at various stages during the item review and revision process.

Two highlights of the test design process are the item review committee and the norming sessions. I organized an item review committee during the test design process to review a draft version of the test. The test consisted of draft items as required by the test specifications. The review committee consisted of language assessment experts from a language testing centre in an Australian university

and visiting testing experts from China. The committee reviewed the test draft and offered feedback in relation to the test construct, task content, task formats, language in the task prompts, time allowance for test-takers, level of test-taker proficiency and other operational concerns in testing. Since there was a mixture of Chinese and non-Chinese testing experts in the committee, the committee was able to provide comments both from a target language perspective and a more general testing perspective. Suggestions for revision from the committee were taken up and incorporated into the test draft.

The two norming procedures were conducted towards the final stage of the test development process. Despite multiple consultations with the four groups of informants (Section 4.1.1.2) in meetings and interviews, I recognized the need to gather more solid quantitative evidence to validate the test tasks before they were finalized. The norming procedures included the administration of norming questionnaires to two groups of L1-Chinese participants and follow-up interviews with the participants to refine the items.

4.1.4 Data analysis

4.1.4.1 TBNA data analysis

The data analysis in the TBNA followed the thematic analysis method as detailed in May et al. (2020) and Pill (2016), a process that derives codes and categories directly from data without relying on *a priori* labels. I first coded both interview and written communication data, organized coded extracts by codes, analysed codes into categories and finally combined categories under super-categories. An example is the data of Alexa recounting a story of her trying to criticize her colleague Ling Mei (pseudonym) at work. Though Alexa perceived her remark as constructive feedback, it somehow damaged their professional relationship. This story was considered one extract out of 236 extracts the entire dataset produced. The process of coding extracts started with locating similarities between them. Compared with Alexa's story, Dave shared a similar story of how criticizing colleagues was challenging for him even though he thought he had never offended anyone. Though being a separate extract, David's story overlapped greatly with Alexa's in content. Henceforth, a code covering both extracts emerged, which was *criticize colleagues*. Out of 236 extracts, 153 codes were identified in this process. I then searched for similarities among codes and coalesced relevant codes into categories. Code *criticize colleagues* and code *criticize friends* were subsumed under the category *criticizing*. Throughout this process, 153 codes were categorized into 49 categories. I finally sought relatedness within categories. For example, the categories *criticizing* and *apologizing* both indicated a concern

about implementing social actions whereas categories *collectivism* and *social hierarchy* oriented more strongly to sociopragmatic issues in L2 Chinese. Out of 49 categories, seven super-categories emerged. In some existing thematic analysis studies super-categories were termed themes. Here, however, the term *super-category* is used in lieu of *theme* because super-category offers a better hierarchical representation of the relationship between super-category and category.

Due to the highly reductive nature of thematic analysis, I conducted a second audit trail with two different L1-Chinese speakers and an applied linguist specializing in qualitative data analysis. All three auditors reviewed the thematic analysis process and questioned me whenever they found anything problematic (e.g., the coding of raw data, the categorizing of categories and the definitions of super-categories). I also implemented a second member checking with all participants by sending out individual reports detailing how their data were analysed. The reports offered information on what super-categories, categories, codes and coded extracts I had extracted from their raw data. This process gave participants the opportunity to express disagreement and actively participate in data analysis. Finally, a second coder coded one interview transcript using the finalized coding scheme and their percentage of exact agreement with me was 86.51 %, which suggested satisfactory intercoder reliability.

4.1.4.2 Test design data analysis

The test design process generated qualitative data from the item editing and moderation process and quantitative data from the norming questionnaires. Informants' feedback for item revision was descriptive in nature and was directly applied to the items when the items were modified. In terms of the norming questionnaire data, descriptive statistics were analysed to evaluate the patterns in norming participants' responses. Details of the questionnaire results are explained in Section 4.2.4.

4.2 Results and initial discussion of study one

Section 4.2 reports the findings from the TBNA and details how test tasks were developed from the results of the TBNA. Section 4.2.1 explains the coding results of the thematic analysis in the TBNA. Section 4.2.2 illustrates how the test specifications were developed from the TBNA thematic analysis results. Section 4.2.3 shows how test tasks were generated from the TBNA and the test specifications while Section 4.2.4 explicates the process of revising and finalizing the test tasks. Section 4.2.5 presents the finalized IC test that was used in this

book. Parts of the findings in the TBNA were presented at the 2019 American Association for Applied Linguistics (AAAL) conference (Dai & Roever, 2019a). Extending findings from this TBNA, Dai (2023a) develops an IC needs model to conceptualize the IC learning needs of L2 speakers. Dai (2023a) also provides detailed lists of L2-Chinese learning goals – a pedagogical toolkit for Chinese language teachers and researchers to systematically incorporate IC in their task-based language teaching, language curricula and language programs.

4.2.1 TBNA results

Though the thematic analysis described in Section 4.1.4.1 followed a bottom-up approach, the reporting of results starts top-down to improve clarity. Figure 4 lists the 7 super-categories identified, ranked by the number of coded extracts. *Social actions* ranked first followed by *sociopragmatic issues* and *pragmalinguistic issues*. *Interactional features, content knowledge, linguistic issues* and *multimodal cues* had much less mention compared to the first three. All the identified seven super-categories shaped the design of IC tasks in the book. Here in Section 4.2.1, I focus on detailing what each of the super-categories contains while in Section 4.2.2 I will explain how content in each of the super-categories contributes to the test specifications.

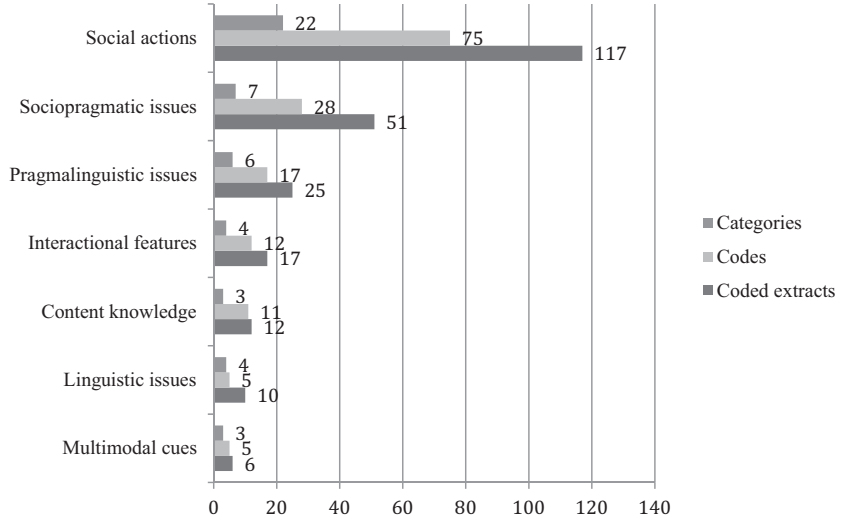


Figure 4 TBNA super-categories

4.2.1.1 Social actions

Figure 5 offers a breakdown of the largest super-category *social actions*, comprising 22 categories, 75 codes and 117 coded extracts. All 22 categories listed in the chart were ranked by the number of coded extracts. Categories such as *apologizing* and *requesting* adopted labels used in traditional speech act research (see Section 2.1.2). Other categories such as *commenting* and *interrupting* are not generally researched in speech act theories but nevertheless fulfil meaningful social actions. It is unsurprising that the number of coded extracts for social actions accounted for nearly half of the total amount of coded extracts produced in the TBNA. As sociolinguists and conversation analysts have long observed, spoken language in communication is inherently pragmatic and performs social actions (Duranti, 1997; Sacks, 1984, 1992). It is the organization and deployment of social actions that L2 speakers need to acquire in order to communicate effectively. Social action is the starting point towards an explication of what causes breakdown in interaction. Understandably it is also the first thing that the TBNA participants oriented to, contributing to its large number of mentions.

If we revisit the definition of affiliation and disaffiliation discussed in Section 2.4.4 and further analyse social action categories by grouping them according to affiliative and disaffiliative actions (Drew et al., 2006), we can see the number of extracts for disaffiliative actions more than doubles the one for affiliative actions. Figure 6 presents three clusters of social actions with disaffiliative actions covering 56 coded extracts and affiliative actions 21 coded extracts. If we unpack the neutral cluster and inspect the 40 coded extracts individually, we can see that though participants reported such incidents in a neutral manner, these incidents can become highly problematic when they are of a disaffiliative nature. Take the code *comment on government officials/political figures* reported by Song Mu in *commenting* for example. Though the action *commenting* is not as polarized as actions such as *criticizing* or *praising* at first glance, it is reasonable to speculate that an inappropriate negative comment would threaten social harmony more gravely than an inappropriate positive comment. Song Mu narrated a story of how she was made to feel uncomfortable when her L2-Chinese friend would not stop discussing Chairman Mao (Mao Zedong) in Chinese with her in a public venue, which drew alarmed looks from onlookers. Though her friends' comments on Chairman Mao were not overtly negative, she found discussing political leaders, especially commenting on historic figures like Chairman Mao, highly objectionable, considering their locale and their degree of closeness as friends. It is reasonable to speculate that had her L2-Chinese friend voiced more distinctive negative comments on Chairman Mao, Song Mu would have been made to feel even more disconcerted.

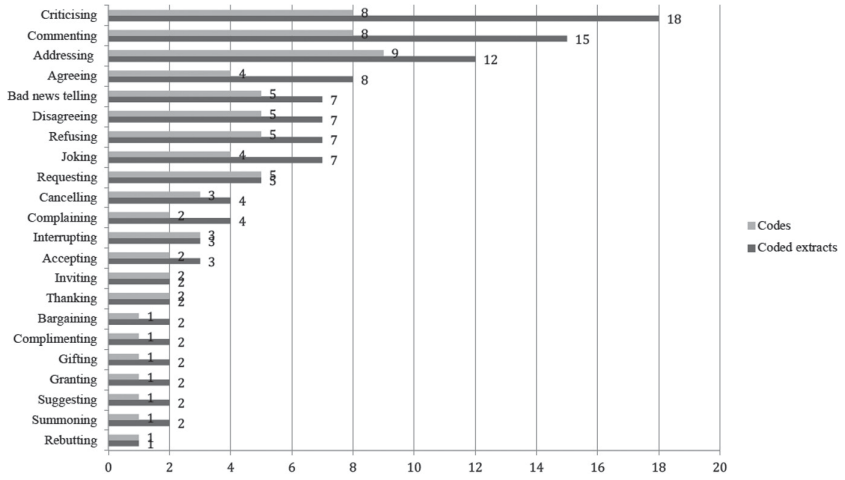


Figure 5 *TBNA social action super-category*

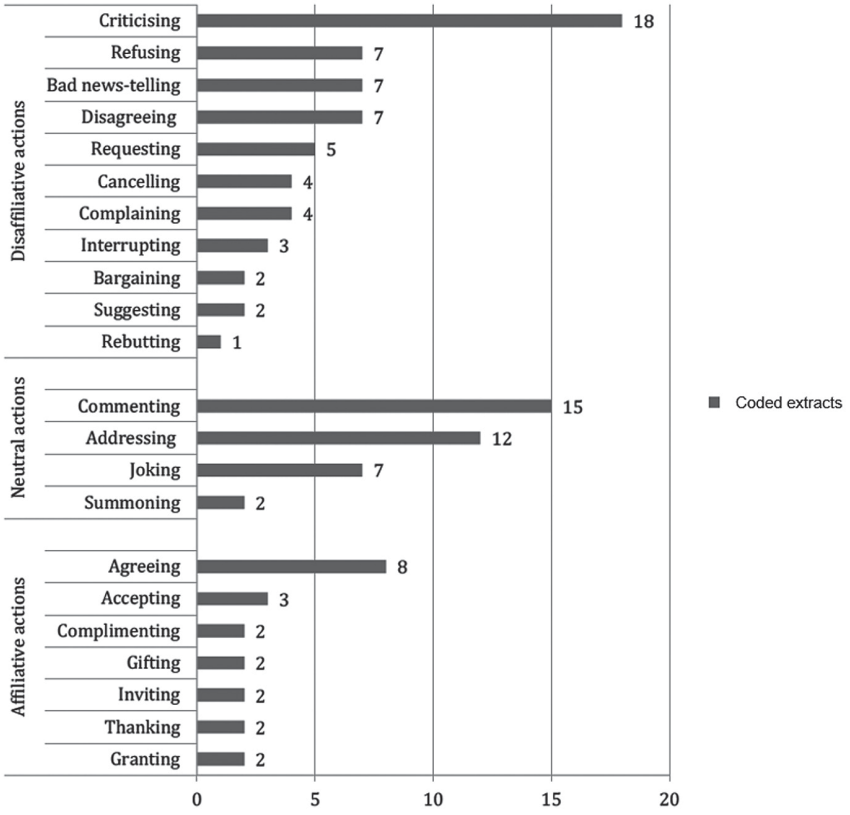


Figure 6 Dis/affiliation in TBNA social actions

Therefore, when taking into account the potential issues that could arise when neutral actions turned disaffiliative, implementing disaffiliative social actions appropriately becomes a paramount concern for L2-Chinese speakers. Based on the high ranking of social actions and the focus on the management of disaffiliative social actions, the IC test developed in this book, therefore, prioritized L2-Chinese speakers' ability to *launch disaffiliative social actions*. The learning and assessment of disaffiliation management warrants more attention as their uncooperative and discordant nature can greatly threaten social harmony and solidarity (Clayman, 2002). The focus on disaffiliation is reflected in the test specifications in Section 4.2.2.

4.2.1.2 Sociopragmatic and pragmalinguistic issues

Sociopragmatic and *pragmalinguistic issues*, on the other hand, were mentioned less often compared to *social actions* as they required the participants and me to probe into critical incidents beyond the surface level. Zhou Wu, an L1-Chinese teacher teaching intermediate Chinese who participated in the TBNA, recounted a story about how he was offended by his student jokingly addressing him as 'ge' (similar to 'bro' in English). Both Zhou Wu and I could immediately tell that the social action at issue was *addressing*. However, what was less clear was what deeper social, cultural, and linguistic factors contributed to this interactional mishap. The H-S interview allowed Zhou Wu and I to delve into the incident, which produced extracts that were later coded under the *sociopragmatic* and *pragmalinguistic* super-categories, such as the code *politeness markers at school* in the category *social hierarchy and language use* in the super-category *pragmalinguistic issues*. Though orthodox conversation analysts resist the evocation of social and cultural artifacts in their analysis of talk (see Section 2.1.4 for a discussion on this), growing research shows that the realization of social actions is cultural-specific and requires appropriate mapping between sociopragmatic and pragmalinguistic knowledge (Golato, 2002; Huth, 2006; Huth & Taleghani-Nikazm, 2006). This is also aligned with the unified pragmatic model on interaction adopted in this book where different schools of pragmatics theories are drawn on to generate an operationalizable philosophical basis for the assessment of IC (see Section 2.1.5). It is therefore unsurprising that most social actions reported in this study had some concomitant sociopragmatic and pragmalinguistic issues, with the former being the second largest super-category (7 categories, 28 codes and 51 coded extracts for *sociopragmatic* in Figure 7) and the latter being the third largest super-category (6 categories, 17 codes and 25 coded extracts for *pragmalinguistic* in Figure 8). How these two super-categories shaped the development of the IC tasks is explored in Section 4.2.2.

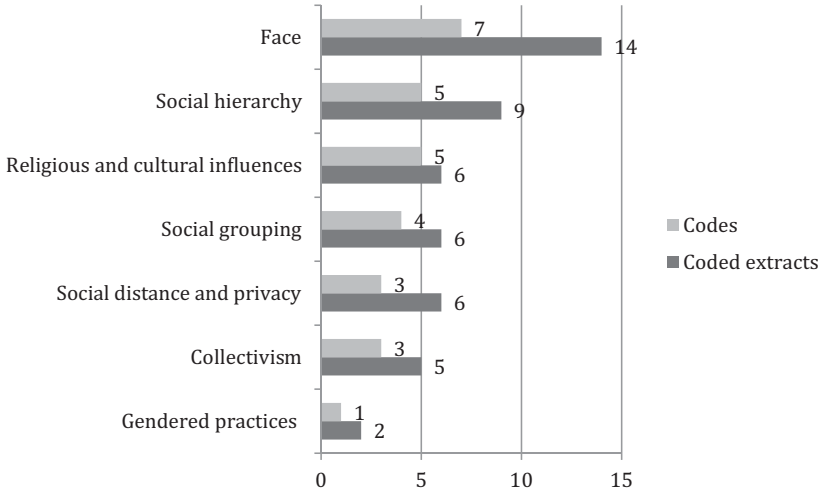


Figure 7 TBNA sociopragmatic super-category

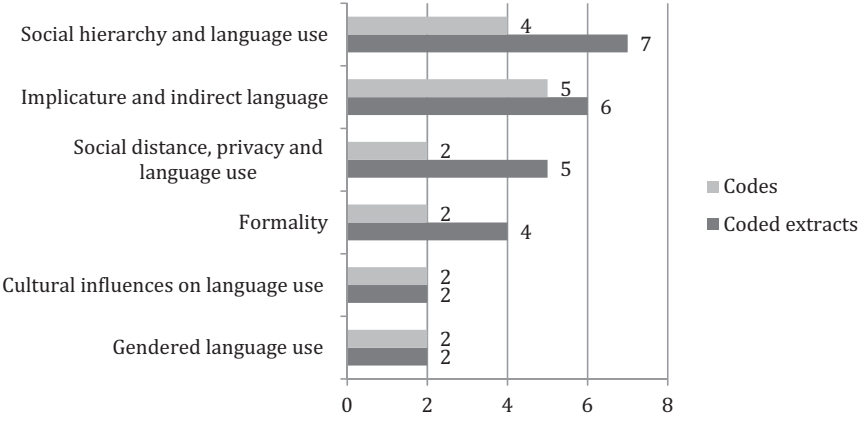


Figure 8 TBNA pragmalinguistic super-category

4.2.1.3 Interactional features and content knowledge

Figure 9 and Figure 10 offer details on the super-categories *interactional features* and *content knowledge* respectively. Though interactional features are usually not perceived by the conscious mind and are difficult to report without CA analysis (Heritage, 1990), it is worth noting that 17 coded extracts in the TBNA were related to interactional patterns and methods, out of which four categories emerged. Applied linguist participants contributed largely to the mentioning of such issues (e.g., *long preambles in storytelling* by Mike in *topic development*) but linguistically untrained participants interestingly also reported issues that are germane to interactional features (e.g., *chatting someone up* by Crissy in *opening*).

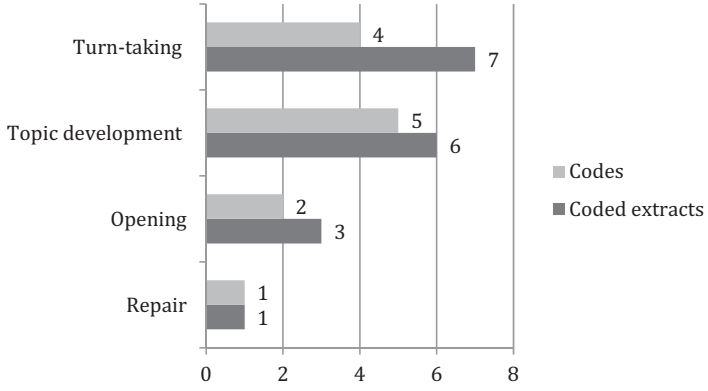


Figure 9 TBNA interactional features super-category

The extracts in *content knowledge*, on the other hand, were mostly contributed by Ella who produced 8 out of the 12 coded extracts in the entire super-category. Representative codes include *talking to real-estate agents* by Dan and *discussing directions with taxi drivers* by Ella, both of whom were beginner L2 speakers. It is telling that setting up life and getting around in China, albeit challenging, quickly became a non-issue for intermediate and advanced L2 speakers, who were more concerned with pragmatic infelicities when the topic for discussion was interactional challenges. The test specifications in Section 4.2.2 took into account participants' reporting of these two super-categories.

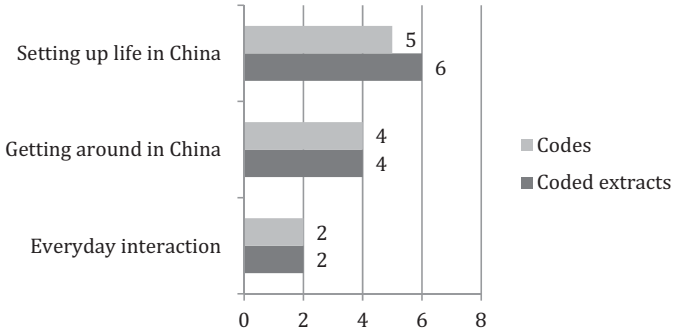


Figure 10 TBNA content knowledge super-category

4.2.1.4 Linguistic issues and multimodal cues

The last two super-categories, *linguistic issues* and *multimodal cues* are presented in Figure 11 and Figure 12. Within *linguistic issues*, *tonality* had the highest number of reports. This is not unexpected given that research in phonetics has shown that L2 speakers in general struggle with tones in the new language, whether their first language is a tonal language or not (Hao, 2012; So & Best, 2010). It is revealing that *grammar* and *vocabulary* hardly received any attention even though they are traditionally the areas of focus in Chinese as a second/foreign language teaching and research (Chu, 1990; Linnell, 2001; Ma et al., 2017). Pragmatic competence and IC, on the other hand, though found to be of crucial importance in this TBNA when it comes to real-world interaction, are still vastly underrepresented and under-researched in the current Chinese as a second/foreign language landscape (Ke, 2012; Li, 2013; Zhao, 2008).

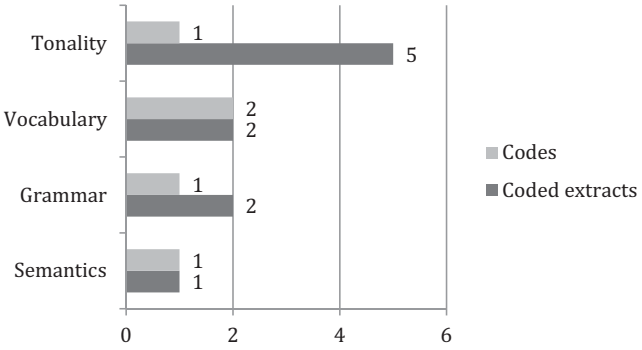


Figure 11 TBNA linguistic issues super-category

Multimodal cues received the least attention of all super-categories. It is understandable that inappropriate *criticizing* can cause more immediate and tangible damage than, for example, inappropriate *eye contact*. However, multimodal cues such as *facial expression* are cultural-specific (Camras et al., 2006; Jack et al., 2012) and can lead to misunderstanding when L2 speakers are unfamiliar with the non-linguistic interactional norms in the L2 community. Research in L2-Chinese multimodal cues is scant though Orton (2014) noted that L1-Chinese teachers in Chinese-as-a-second-language teaching contexts in China highlighted the importance of facial expression management when assessing Chinese learners' speaking performances. Such an emphasis was not placed by L1-Chinese and L2-Chinese teachers in Australia in Chinese-as-a-foreign-language contexts in Orton's study. This can be interpreted as incipient evidence that multimodal cues, though constituting an integral component of effective Chinese interaction in China, tends to be neglected by Chinese teachers not residing in the target community. More research is needed in this direction to better our understanding of the interactional import of multimodal cues in L2-Chinese interaction.

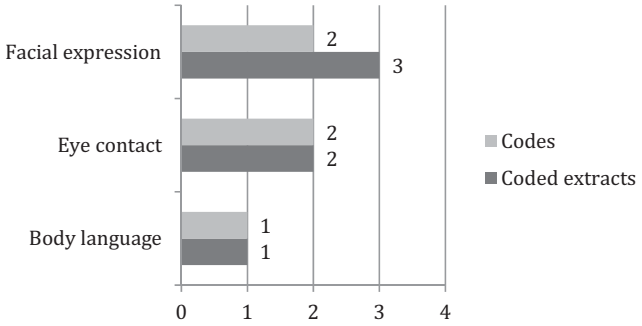


Figure 12 TBNA multimodal cues super-category

4.2.2 The test specifications

The insight from the TBNA as reported in Section 4.2.1 served as the starting point for the design of assessment tasks of IC. This section explains how findings from the TBNA translated to the test specifications for the IC test in this book, following each of the super-categories reported previously.

The *social actions* super-category in Section 4.2.1.1 points to the observation that disaffiliative social actions are what L2-Chinese speakers find most challenging. Effective management of disaffiliative situations is also crucial to interactional success, as disaffiliation, if unchecked, can cause strain on social solidarity and interpersonal relations. In view of this, the IC test in this book focuses primarily on interactional tasks where test-takers need to demonstrate their IC in managing disaffiliation. Each task in the test requires test-takers to handle a particular disaffiliative social action and test-takers will be assessed on their ability to undertake such endeavours successfully.

The discussion on the *sociopragmatic* and *pragmalinguistic* super-categories in Section 4.2.1.2 suggests a range of contextual sociocultural factors that influence the launching of social actions and the linguistic choice in the realization of social actions. In order to control for such contextual variables in an assessment setting, I adopted Brown and Levinson's politeness theory to regulate the contexts where social actions take place in the IC test. For the three contextual variables in politeness theory, I decided to keep the social distance and rank of imposition variables constant, while varying the power (P) variable. The rationale for varying the power variable is that power or social hierarchy is a recurring issue in the *sociopragmatic* and *pragmalinguistic* super-categories in the TBNA, ranking

second in the *sociopragmatic* category and first in the *pragmalinguistic* category in terms of saliency (see Figure 7 and Figure 8). The example of misusing address terms shared by Zhou Wu in Section 4.2.1.2 highlights that power is an inherent factor mediating interaction and if mismanaged, can cause severe disruption to social harmony. The IC test tasks will therefore require test-takers to demonstrate the ability to interact successfully in social situations where they have less power than their interlocutor (P−), the same amount of power as their interlocutor (P=), or more power than their interlocutor (P+). The other two variables, social distance and rank of imposition, need to stay constant so as to ensure the number of tasks is manageable in a testing setting. In particular, I decided to set the social distance to *medium to large* as research shows that it is more challenging to interact when the speaker is less familiar with their interlocutor (Wolfson, 1990). For the rank of imposition variable, I chose to set it to *high* so as to make the test tasks more demanding for test-takers. The rationale for adjusting the two constant variables to make the test tasks harder is that pragmatic/IC assessment tasks have frequently proven to be under-challenging for the intended test-taker population (Roever, 2018). More demanding IC tasks can better differentiate test-takers at different IC levels.

Having identified the social actions to target and having controlled the contextual variables, I moved on to other TBNA super-categories to flesh out the test specifications and decide on the task types for the test. The *interactional features* super-category emphasizes the importance of incorporating interactive tasks in the test. The measurement of IC also requires a high degree of interaction in the test tasks so as to model real-world interaction. To address this issue, I chose two sites of CMC – asynchronous voice messaging and synchronous video chat – as bases for the design of tasks. Voice messaging was further divided into two task types: first-pair-part voice messaging and second-pair-part voice messaging. By using different methods of testing, I aimed at reducing the method effect so as to offer a more accurate depiction of test-taker IC. The three test methods also represent different degrees of interactiveness. First-pair-part voice messaging is the least interactive method as it requires test-takers to only initiate voice messages. Second-pair-part voice messaging is more interactive since test-takers need to listen first and then speak but the interaction is still asynchronous. Video chat is the most interactive method as test-takers need to orient to contingencies in talk and produce sequentially meaningful speech online. Both voice messaging and video chat will be conducted on WeChat, a mobile phone-based application that enjoys wide popularity in Chinese-speaking communities. Using a realistic phone-based application for test delivery can enhance test authenticity and reinforce the

link between assessment and positive washback on everyday practices. The inclusion of different spoken CMC modes also addresses the research gaps in CMC literature as identified in Section 2.2.

Finally, the *linguistic issues* super-category, comprised of mostly linguistic competence (LC) indicators, was not foregrounded in the IC test specifications due to the relatively less focus it received from the TBNA participants. However, challenges with linguistic devices are reflected in the launch of social actions in the test. Since the focus of the IC test is on interaction rather than linguistic knowledge, a lesser focus on purely assessing LC is justified in this case. The assessment of *multimodal cues*, as reported in the last super-category in the TBNA, warrants consideration as the multimodality of interaction can pose challenges to L2 speakers. However, given that multimodal cues were rarely mentioned by the TBNA participants and logistic considerations of assessment, non-linguistic features and multimodal interaction were not included in the test specifications.

Having translated the TBNA findings in Section 4.2.1 to considerations of test development, in Table 12 below I present the resultant test specifications following the checklist in Alderson et al. (1995). Some of the checklist items are not defined at this stage of test development but will be specified when the rating scale design takes place in Chapter 5. In such cases, references to book chapters and sections are provided for ease of navigation.

Table 13 below illustrates the task shell of the IC test, showing how the test tasks are mapped onto test methods, sub-target language use (TLU) domains and the power (P) variable. I ensured there was a task for each of the three delivery methods and each of the three P conditions in each of the three TLU domains. The next step in test design is to generate items that match the item specifications and task shell.

Table 12 Test specifications

Preliminary test specifications	
Purpose of the test	To assess the IC of L2-Chinese speakers who intend to migrate to China and who need to demonstrate their ability to interact successfully in work, study and everyday interaction
Description of test-takers	Speakers of Chinese as a second/additional language
Test level	Suitable for test-takers that have basic Chinese proficiency (at or above HSK 3)
Construct	Interactional competence (details regarding how the construct is defined can be found in Chapter 5)
Number of tasks	Nine (how this is arrived at is explained in Table 13)
Time for each task	Depending on test-takers' needs but in general, each task should be completed within five minutes (based on pre-pilots and trials to be detailed in Section 4.2.4)
Weighting for each task	Same weighting is given to each task
TLU domain	Interaction in everyday, study and workplace settings
Language skills to be tested	IC/talking (different from LC/speaking as discussed in Section 2.3)
Language elements to be tested	Indicators of IC (details regarding how the indicators are defined and selected can be found in Chapter 5)
Test tasks	Role-play tasks
Test methods	1st pair-part voice messaging, 2nd pair-part voice messaging and live video chat
Rubrics	Analytic rating scale (details in Chapter 5)
Criteria for marking	Test-taker performance needs to match most of the descriptors in a particular band level in the rating scale to qualify a score at that level
Description of typical performance at each level	Descriptors at each band level in the analytic rating scale developed in Chapter 5
Description of what candidates at each level can do in the real world	See the definition of what each band in the analytic scale corresponds to real-life competence in Table 29 in Chapter 5
Samples of students' performance on tasks	Test-taker exemplars developed in Chapter 5

Table 13 The test shell

		1st pair part voice messages	2nd pair part voice messages	Video chat
Life	**P+**			X
	P=		X	
	P-	X		
Study	**P+**	X		
	P=			X
	P-		X	
Work	**P+**		X	
	P=	X		
	P-			X

Note. P stands for the power variable in politeness theory

4.2.3 Generating draft items

Having identified the test specifications and task shells in Section 4.2.2, in this section I focus on how I drafted sample items. Going back to the TBNA results, the rich details in the critical incidents elicited by S-H interviews allowed the critical incidents to be easily translated into authentic and meaningful communicative language tasks as advocated in Long (2016).

Here a sample task focusing on the code *how to cancel plans you already made with Chinese friends* under the category *cancelling* is presented. First, I reviewed the coded critical incidents in this category. One coded extract was contributed by the TBNA L2-Chinese participant Dan (see Table 9 for Dan's profile), who recounted a story of how he once got into an argument with his L1-Chinese friend Li Ming (pseudonym). Li Ming made a prior arrangement with some of his friends to go to a bar that night when he met Daniel for dinner. Daniel tried hard to persuade Li Ming to give up that plan and go clubbing with him instead. Though clearly tempted by Daniel's proposal, Li Ming hesitated, leading to Daniel becoming more direct in his persuasion. To Daniel's surprise, Li Ming appeared defensive and said: 'no I can't just cancel my plan like that. You don't understand how it works because you are not Chinese.' An altercation followed and Daniel still felt quite puzzled by Li Ming's behaviour when participating in the TBNA.

Reviewing the interview transcripts, I noticed that I engaged in a discussion with Daniel on this critical incident and offered some possible explanations for the communication breakdown. Li Ming's reaction could be triggered by the concern that cancelling plans last minute was perceived, at least according to

Li Ming, as greatly face-threatening in Chinese culture. Daniel reflected that though cancelling was also disaffiliative in his home country Australia, it is much more common compared to in China as Australians reschedule or cancel plans with friends all the time with no serious repercussions. Another possible explanation is that the plan Daniel wanted Li Ming to get out of involved a group of Li Ming's friends. Due to the perceived collectivistic culture in Chinese society (Hofstede Insight, 2021), Li Ming might have found it harder to extract himself from a commitment that was made to a group of people. The analysis offered here is admitted incomprehensive and rudimentary, based on assumptions and generalizations of cultures. This critical incident, however, does lend itself to the development of an authentic language task that L2 Chinese speakers find challenging.

Relying on the rich details in this critical incident from the H-S interview, I developed a sample IC task where L2-Chinese speakers need to cancel a plan they have made with a group of close friends and by extrapolation with a group of colleagues in a workplace setting. This would be a P= task in either everyday interaction or workplace language use domains. This task could be implemented either through voice messages or video chats, though voice messages would be a more plausible method given the relatively low stakes of cancelling in this context.

Following a similar process and drawing on other critical incidents reported in the TBNA, I drafted an item bank consisting of various items that fitted the item shells in Table 13. Although the test only required nine items, more items were designed in case some items were deemed unsuitable and needed to be deleted during item trials and pre-pilots.

4.2.4 Revising the draft items

As discussed previously, I wrote up a pool of draft items based on findings from the TBNA. The items were then presented to the four groups of participants for item editing and moderation: L1-Chinese speakers, L2-Chinese speakers, Chinese language teachers, applied linguists and testing specialists (see Section 4.1.1.2). It should be reiterated that the four groups of participants for item review and moderation in Section 4.1.1.2 were different from the three groups of TBNA participants in Section 4.1.1.1. The TBNA participants helped to generate the TBNA findings which informed my writing of the draft items while the item review and moderation participants in Section 4.1.1.2 assisted me to revise the draft items. Select members from the four groups of item review and moderation participants also trialled the items to assess their functioning (Alderson et al.,

1995). Details regarding the backgrounds of the item review and moderation participants, the use of an item review committee and the iterative process of item trials and pre-pilots can be found in Section 4.1.1.2 and Section 4.1.3.2. After the draft items went through the item review and moderation process, they were assembled into a draft nine-item IC test, which then underwent two rounds of norming with L1-Chinese speakers (see Section 4.1.1.2, 4.1.3.2 and 4.1.4.2 for more details).

Section 4.1.2.2 explains that the norming questionnaire consisted of a DCT question that asked respondents to write down what they would say in that scenario and five Likert scale items that quantitatively validated the items. DCT responses confirmed that most L1-Chinese speakers oriented to the social actions as intended by the researcher, indicating that the test tasks functioned well in eliciting the intended interactional behaviour. Questions two to six in the questionnaire were concerned with the appropriateness of the test tasks and the validation of the contextual variables. These five questions are Likert scale questions that asked respondents to choose a number from 0 to 10 to indicate their opinion. Here I present and discuss the descriptive statistics from the first norming session, which was administered to 33 participants. A complete version of the norming questionnaire in both Chinese and English can be found in Appendix II.

Table 14 represents the norming results for Question two, which investigated how realistic L1 speakers considered the tasks to be. The Likert scale goes from 0 to 10, with 0 being *very unrealistic*, 3 being *not realistic*, 5 being *average*, 7 being *realistic*, and 10 being *very realistic*. The nine items are labelled based on their Power (P) variable and the sub-TLU domains they are in, as illustrated in Table 13. SD stands for standard deviation. Results show that on average the 33 L1 speakers considered the nine task scenarios to be of acceptable levels of realism, which ensured the authenticity of the tasks. There are some possible reasons why on average the norming participants did not assign the tasks extremely high realism scores. First, the tasks were assessment tasks that were abstracted from real-world situations and therefore could never fully realistically replicate the rich situational and contextual information of real-world activities. All norming participants were everyday-life people living in China with no prior training in language teaching or assessment. They, therefore, did not have the insight into the constraints of language tasks as test developers do. Second, the tasks were designed to be highly pragmatically challenging (Roever, 2018) to elicit a range of IC abilities for assessment purposes. This might have dampened norming participants' perception of the realism of the tasks, though the tasks were clearly still perceived to be more realistic than not. It was a balancing act

I had to manage carefully in item development and revision: on one hand, a good IC test needs to have items that are difficult enough to discriminate between test-takers of higher IC levels, which has been an issue noted in previous IC assessment research where the test tasks were on the easier side for the test-taker cohorts (Ikeda, 2017; Youn, 2013). On the other hand, the test tasks cannot be so taxing to the extent that they are no longer considered realistic or likely to happen in real life. The realism ratings achieved in the norming session were considered satisfactory for the purpose and context of the test.

Table 14 Realism of the test items

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Item	**N**	**Min**	**Max**	**Mean**	**SD**
Life P+	27	3	10	6.41	1.69
Life P=	31	4	10	7.06	1.73
Life P-	31	4	10	6.23	2.25
Study P+	33	3	10	6.24	1.66
Study P=	31	3	10	6.65	1.89
Study P-	32	3	10	6.59	1.54
Work P+	31	3	10	6.26	1.65
Work P=	32	3	10	6.88	1.60
Work P-	31	3	10	5.71	1.92

Table 15 explains the results for Question three, which is how appropriate it seemed to the norming L1-Chinese participants to use the designated interactional method (1st pair-part voice messaging, 2nd pair-part voice messaging or video chat) to approach the interactional scenarios in the respective test tasks. The abbreviations of the interactional methods are included in Table 15 for ease of reference, with 1st PP referring to 1st part-part voice messaging, 2nd PP referring to 2nd pair-part voice messaging and live referring to video chat. On a scale from 0 to 10, 0 means the interactional method is *very inappropriate* and 10 means it is *very appropriate*. Similar to the responses to Question two, most respondents found the matching between interactional methods and task scenarios *above average* in terms of appropriateness. Possible explanations for not achieving extremely high appropriateness scores are similar to the ones offered

for the results of the previous realism item. Due to the abstracted nature of the assessment tasks and the demanding nature of the tasks in this IC test, norming participants who were unfamiliar with task development might hesitate to award very high appropriateness scores to the tasks, as they would be comparing language tasks with real-world activities that are always much more nuanced and complex than what language tasks can capture. It was encouraging to see that even norming participants who were untrained in language testing still considered the tasks in general above average in terms of task delivery methods, which enhanced the extrapolative strength of the tasks to real-world activities.

Table 15 Method appropriateness of the test items

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Item	**N**	**Min**	**Max**	**Mean**	**SD**
Life P+_video	29	3	10	6.55	1.94
Life P=_2nd PP	32	0	9	5.19	2.56
Life P-_1st PP	32	3	10	6.25	2.10
Study P+_1st PP	33	3	10	6.55	2.09
Study P=_video	32	3	9	6.09	1.59
Study P-_2nd PP	33	2	10	5.70	2.22
Work P+_2nd PP	29	3	10	5.14	2.31
Work P=_1st PP	32	3	10	6.97	1.58
Work P-_video	29	3	9	5.45	1.88

Table 16 focuses on the rank of imposition variable, which as discussed in the test specifications in Section 4.2.2, was made to stay constant in this IC test. It was also considered ideal if the imposition variable was high so that the test tasks could be more challenging for L2-Chinese speakers, considering existing L2 IC tests are not sufficiently covering advanced test-takers' abilities (Roever, 2018). On the scale from 0 to 10, 0 means *very easy to implement* while 10 means *very difficult to implement* in terms of imposition. Results show that respondents found most items to be of similar difficulty, floating around 5 out of 10. This level of difficulty is desirable as an *average* difficulty for L1-Chinese speakers would be sufficiently difficult for L2 speakers, especially when the test was designed

to cover a wide range of proficiency as explained in the test specifications in Table 12. The only item that was noticeably more difficult than others is the life P= item. This item clearly needed revision and will be discussed in more detail shortly.

Table 16 Rank of imposition of the test items

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Item	**N**	**Min**	**Max**	**Mean**	**SD**
Life P+	25	1	9	4.64	1.85
Life P=	33	5	10	7.55*	1.50
Life P-	31	3	10	5.81	1.74
Study P+	25	1	10	5.48	2.08
Study P=	31	3	8	5.42	1.54
Study P-	31	3	10	5.68	1.92
Work P+	27	3	10	5.19	1.73
Work P=	30	3	8	4.80	1.58
Work P-	28	1	10	5.64	2.04

Table 17 illustrates the results of the rating for the social distance variable, which was also supposed to be constant. The results are in general satisfactory as most respondents gave scores within the 5–6 range, indicating *average* familiarity between the test-taker and the interlocutor. This conforms with the requirement in the test specifications as I wanted the social distance to be around medium, which theoretically is considered to be more challenging to manage (Wolfson, 1990). The life P= item, similar to the one in Table 16, also noticeably departed from other items since respondents found the relationship between the test-taker and the interlocutor to be closer than other items. The study P= item also warrants some closer inspection as the average is also slightly higher than other items.

Table 17 Social distance of the test items

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Item	**N**	**Min**	**Max**	**Mean**	**SD**
Life P+	30	3	10	6.37	2.08
Life P=	33	5	10	8.48*	1.52
Life P-	31	3	10	5.55	1.67
Study P+	33	3	10	6.82	1.78
Study P=	33	5	10	7.36*	1.32
Study P-	31	3	10	6.03	1.38
Work P+	33	2	10	5.61	1.92
Work P=	33	3	10	5.85	1.40
Work P-	31	2	7	5.32	1.38

Finally, for the power variable, Table 18 demonstrates that the L1-Chinese speaker respondents corroborated my judgement. Question six, the question that enquires about the power variable in the norming questionnaire, was phrased in a way that asked the respondent to rate the interlocutor's power compared to the one of the test-taker. Therefore, a higher score in Table 18 corresponds to a lower power rank in the test-taker. The naming of the items, on the other hand, comes from the test-taker perspective so items with P- indicate that the test-takers are of lower power ranks than the interlocutor, which, if the items function as expected, should receive higher scores in the table. Results in Table 18 confirm my expectation as items with P- had consistently higher means (6.96, 9.24 and 7.09), demonstrating that questionnaire respondents considered test-takers to have less power than their interlocutors. P= items had consistent average power ratings (5.18, 5.18 and 5.39) which indicates that respondents viewed test-takers as having similar power to their interlocutors. P+ items also performed well, as they consistently gained lower mean scores (3.52, 3.16 and 3.82), showing that respondents thought of the interlocutors as having less power than the test-takers.

Table 18 Power of the test items

0	1	2	3	4	5	6	7	8	9	10
Much lower			Lower		Same		Higher			Much higher

Item	N	Min	Max	Mean	SD
Life P+	29	0	7	3.52	1.35
Life P=	33	5	7	5.18	0.53
Life P-	24	5	8	6.96	0.75
Study P+	31	0	6	3.16	1.51
Study P=	33	5	10	5.18	0.88
Study P-	33	4	10	8.24	1.52
Work P+	33	0	8	3.82	1.67
Work P=	33	3	7	5.39	0.86
Work P-	32	3	10	7.09	1.53

Having examined all nine items for their performance on the norming questionnaire, it is clear that most items functioned well except for item life P=, which has higher than expected imposition and social distance ratings. Referring back to the task shell in Table 13, we can see this is a second-pair-part voice messaging item. The task prompt for this item is reproduced below in English translation.

> Wang Si is your friend. You two are similar in age. She went traveling overseas for three months and asked you to look after her dog, Dou Dou, whom she had had for eight years. Last week you took Dou Dou out for a stroll, met a friend and started a chat. You were distracted by the chat and let go of the leash. When you realized it, Dou Dou was already gone. You have looked for Dou Dou for three days but still haven't found him. Just now Wang Si sent you a voice message. After listening to it you decide to reply with a voice message.

> Wang Si's voice message: 'Hey, my friend, how are you doing? Will you be home this afternoon? I just realized Dou Dou is due for his check-up today and I have arranged for a friend to pick him up this afternoon and take him to the vets. Could you give Dou Dou to my friend when he arrives? Thanks!'

After discussing this item with some of the norming participants, I realized in the Chinese context a friend would only entrust the responsibility of looking after their pet to someone who is very close to them, which explains the higher familiarity score for social distance. Losing a friend's dog is perceived very negatively and is in a sense potentially not redeemable, which makes the imposition score very high as well. I subsequently revised this item to the following scenario, which still places the test-taker in an equal power position to the interlocutor in an everyday interaction setting.

> Wang Si is your neighbour. You two are similar in age. One week ago, she went to another city for some urgent errands and before she left, she asked you to send a parcel for her. You promised her you would send it on the same day. However, you were very busy with work for the following week and completely forgot about the parcel. Just now Wang Si sent you a voice message. After listening to it you decide to reply with a voice message.

> Wang Si's message: 'Hey, I just wanted to check if you have already sent off my parcel? Actually, it is my friend's birthday present in the parcel. My friend had their birthday yesterday and strangely they said they didn't receive my gift. I found it quite weird because the parcel should have arrived after one week.'

After revising this life P= item and making some adjustments to other items based on the first round of norming, I conducted a second round of norming using the same norming questionnaire with a different group of 21 L1-Chinese participants. The results were similar to the first round with the performances of the life P= item and the other revised items more aligned with my expectation and the test specifications.

4.2.5 Finalizing the IC test

After the multi-stage iterative process of item revision, trials and pre-pilots, the IC test was finalized as a nine-item test with each item targeting a disaffiliative social action.I selected the highest-rated disaffiliative social actions in the TBNA (see Figure 6) and designed nine IC speaking test items. 'Criticizing', the top-rated social action in the TBNA, appears twice in two items of different sub-TLU domains whereas each of the other seven social actions appears once in the remaining seven items, offering maximal coverage of the social actions L2-Chinese speakers found most challenging. Table 19 is a revised version of the task shells in Table 13 as Table 19 explains the disaffiliative action each item targets.

Table 19 Task shell populated

		1st pair part voice messages	2nd pair part voice messages	Video chat
Life	**P+**			Criticizing
	P=		Breaking bad news	
	P-	Requesting		
Study	**P+**	Negative commenting		
	P=			Disagreeing
	P-		Refusing	
Work	**P+**		Criticizing	
	P=	Cancelling		
	P-			Complaining

Building on the test specifications in Table 12, here I offer a brief summary of the IC test to summarize the features and structure of the test.

This nine-item IC test was designed to assess the IC of L2-Chinese speakers. The nine items in the test are based on a TBNA on interactional scenarios that L2-Chinese speakers found most challenging. The test is delivered entirely via CMC instead of F2FC to improve test practicality and accessibility. The CMC platform used is a communication mobile phone application called WeChat, which is popular among Chinese-speaking users and communities. Three types of tasks of increasing interactiveness are included in the test: initiating first pair part voice messages, supplying second pair part voice messages and interacting with interlocutors in video chats. The nine tasks span three sub-language use domains: everyday interaction (life), education-related communication (study) and workplace communication (work). Each one of the nine scenarios targets a particular disaffiliative social action, such as breaking bad news to a neighbour or cancelling a plan you have made with a work colleague. The interlocutors in the tasks vary in power, with some having more power than the test-takers' social roles in the tasks, some of the equal power, and some having less power. The other two contextual politeness variables, rank of imposition and social distance (Brown & Levinson, 1987), are held constant. The ratings of politeness variables, the authenticity of task scenarios, and the appropriateness of task delivery methods were validated via both qualitative and quantitative evidence. The items in the test can be found in Appendix III. For each item test-takers received a video prompt containing a Chinese voiceover and still images that illustrate the scenarios. All the images in the video prompts were hand drawn by the researcher. It should be noted the test-taker in the video prompts was drawn

to be gender-neutral so that test-takers of different gender identities can relate to the prompts. English translations of the Chinese voiceover are also provided in Appendix III.

The ordering of the items, from item 1 to item 9, follows the order of item delivery in the main testing session in study three in Chapter 6. The order is based on item difficulty in the pilot test study, which is explained in Chapter 5. The naming of the items follows item order_TLU sub-domain_delievery method_ Power variable_social action. Therefore item 1 is named 'Item 1_study_1st pp_P +_negative commenting', which means this is the first item to be delivered in the IC test, it pertains to the study sub-TLU domain, it requires test-takers to supply a 1st pair-part voice message, the test-taker has more power than the interlocutor in this item, and the intended social action is negative commenting.

One advantageous feature worth highlighting of the IC test is that it was designed across a large domain of language use, comprising effectively three sub-TLU domains of educational, workplace and everyday settings. This was a purposefully made decision for benefits intended for test users. The discussion on the domain description inference in the interpretive argument in Section 3.1.1 explains why a large TLU domain can be desirable as it allows test users to infer test-takers' language ability in a wider range of situations, instead of ability in a very restricted situation such as the speaking skills in giving an oral presentation in a third-year literary criticism class in an undergraduate Arts degree in an Australian university. Considering the advantages in the usability of IC tests with a larger TLU domain, I designed the IC test in this book to include all three sub-language use domains, namely study, work, and everyday interaction. The domains of workplace and everyday interaction are rarely covered in previous IC assessment, which predominantly focuses on interaction in the academic context (Ikeda, 2017; Youn, 2013). However, as demonstrated by the TBNA results, L2-Chinese speakers have genuine interactional needs in everyday interaction (see the *content knowledge* super-category in Section 4.2.1.3) and as they progress in their lengths of residence in the Chinese-speaking communities, they start to encounter professional interactional challenges in the workplace (Wang, 2011). The generation of items across three subdomains was facilitated by the triangulation of needs analysis informants with expertise in different domains of language use (see Section 4.1.1.1). The three sub-language use domains in this IC test can assist relevant stakeholders in developing a better understanding of how ready test-takers are to relocate to Chinese-speaking communities for living, study, and work.

Having said that, I acknowledge that not all potential test-takers would have had experience in all the task scenarios in the three sub-TLU domains

as some test-takers might have never studied or worked in Chinese-speaking communities. This was however not considered an issue given the proposed uses of the test. As stated throughout this chapter and more specifically in the test specifications in Table 12, this IC test was created to assess L2-Chinese speakers' IC if they are to migrate to Chinese-speaking communities and interact verbally in everyday, tertiary study and workplace contexts. In order for test score users to understand test-takers' readiness for language use in this TLU domain, test-takers need to demonstrate the requisite language skills in the domain even though they might not have had experience in some situations in the domain. This is no different from IELTS assessing test-takers' ability to interact in tertiary education settings when the test-takers have never interacted in universities where English is the medium of instruction, or OET assessing overseas trained doctors' or nurses' ability to undertake clinical communication in English-speaking countries when they have never lived or worked in such countries.

It should also be noted that the three sub-TLU domains in the test contain scenarios that are *general* instead of *specialized* in nature. The final items admitted into the test in Appendix III cover interactional scenarios that every test-taker would have had some knowledge of or experience in, if not in their L2 but certainly in their L1. Assessing test-takers' ability to talk with classmates, shop owners, neighbours and colleagues on general everyday-life topics is very different from assessing their ability to demonstrate language skills as a doctor, pilot, or customs officer, where specialized language and content knowledge are expected. Therefore, the IC test in this book is an appropriate instrument to measure test-takers' ability to interact for general purposes in a comprehensive TLU domain.

Another point worth highlighting is that the test is designed to cover content in all the seven super-categories from the TBNA. Although the test centres disaffiliative social actions in the *social actions* super-category, content in other super-categories is also incorporated in the test tasks. Due to the employment of the three contextual variables in politeness theory, items in the test vary in terms of *social hierarchy*, *social grouping*, *social distance* and *formality*, which are some of the key categories in the *sociopragmatic* and *pragmalinguistic* super-categories. The gender of the interlocutors in the nine items is mixed, addressing the *gendered practices* and *gendered language use* categories in the *sociopragmatic* and *pragmalinguistic* super-categories. Since the test is interactive, it captures the categories in the *interactional features* super-category such as *turn-taking* and *opening*. The test offers good coverage of the codes in the *linguistic issues* super-category due to its focus on productive skills. The three live chat items have the capacity to elicit codes in the *multimodal cues* super-category, though

the assessment of multimodal IC is not considered for this book due to logistical restrictions in rating. In summary, the test offers sound coverage of the learning needs of L2-Chinese as identified in the TBNA.

Details regarding how the test was pilot-tested and administered to the main testing population can be found in Chapters 5 and 6 respectively. It should be noted that at this stage the construct of the test is still largely unspecified. Chapter 5 addresses this issue by defining an operational IC test construct for this test through the use of indigenous criteria combined with theories in sociology and philosophy.

Chapter 5 Study two: Pilot test, indigenous criteria, and rating materials

Having developed the IC test based on the TBNA and a methodical test development process in Chapter 4, Chapter 5 explains how the test construct was specified and how the rating materials were developed. The rationale for conducting study two and addressing the research questions of study two is detailed in Section 3.2.2.

The development of the test construct and rating materials started with collecting test-taker sample performance to be used as stimuli. A pilot test with 22 test-takers was conducted and test-taker performances at different ability levels were elicited. The pilot test-taker performances were rated by three pilot raters using a holistic three-step scale to judge their IC. Rasch analysis demonstrated high rater reliability, strong item performance, and preliminary evidence for unidimensionality.

The pilot test-taker performances were used as stimuli to elicit 36 everyday-life domain experts (DEs)' indigenous criteria of IC. Thematic analysis of interview transcripts and written comments from DEs yielded a five-category representation of indigenous IC criteria: (1) conflict management, (2) reasoning skills, (3) solidarity promotion, (4) personal qualities, and (5) social relations. I then thematically coded the DEs' comments within each of the five categories to different ability levels and produced a five-step rating design: (1) exemplary, (2) good, (3) average, (4) development needed, and (5) intervention needed. A third thematic analysis was conducted on DEs' comments within each rating category to identify sub-categories. The upshot of this process was a five-step indigenous IC rating scale of five rating categories with sub-categories and sub-category descriptors.

Due to the a-theoretical nature of DEs' indigenous criteria, the indigenous scale is precise in its measurement of test-taker performance for this particular test and particular L2. To broaden the applicability of the IC scale in this book, I utilized Sequential-Categorial Analysis, combining Conversation Analysis (CA) and Membership Categorization Analysis (MCA), to empirically validate DEs' comments and indigenous criteria through pilot test-taker performance data. This emic validation process allowed the descriptors in the indigenous scale to be theorized via CA and MCA concepts, which apply across different interactional contexts, test tasks, and L2s. The resultant theorized IC scale has five theorized rating categories: (1) disaffiliation control, (2) affiliation

promotion, 3) morality, 4) reasoning, and 5) social role management. This new rating scale offers broader applicability beyond the assessment context of this book and encompasses not only the sequential but also the emotional, moral, logical and categorial dimensions of interaction. Since a rating scale embodies the test construct, the theorized IC scale is theory-generating in its nature as it represents a more holistic conceptualization of IC, which goes beyond the mechanics of interaction. This proposed model of IC is theoretically robust as it is aligned with other holistic models of interpersonal communication such as the original Hymes-ean model of communicative competence and the artistic proofs in Aristotelian rhetoric.

5.1 Methodology of study two

Section 5.1 explains the methodology adopted in study two and Section 5.2 focuses on the results and initial discussions. Highlights of the methodology of study two include the employment of everyday-member DEs' indigenous criteria to specify the IC test construct, the broadening of the IC test construct by drawing on theories and concepts in sociology and philosophy, the validation of rating categories via CA and MCA, and lastly, the consultation with domain expert feedback in the design of steps in the rating scale.

5.1.1 Participants

5.1.1.1 Pilot test test-takers

To examine the functioning of the test and to generate a wide range of performances for DEs to comment on, I administered the test to a pilot test group of 22 test-takers comprising 11 L1-Chinese speakers and 11 L2 speakers of differing Chinese proficiency and varying length of residence in China. The rationale for selecting a diverse group of pilot test test-takers with varying degrees of proficiencies and socialization in the L2 community is to improve the generalizability of the rating scale to be developed. A rating scale based on a heterogeneous test-taker sample implies the scale can capture the large variability in the real-world Chinese speaker population.

The 11 L2-Chinese test-takers came from a range of L1s, as summarized in Table 20.

Table 20 Pilot L2-Chinese test-takers' L1s

L1	N	%
Australian English	5	45.45 %
British English	1	9.09 %
American English	1	9.09 %
German	1	9.09 %
Italian	1	9.09 %
Hebrew	1	9.09 %
Vietnamese	1	9.09 %

Nine out of the 11 L2-Chinese test-takers had taken the HSK Chinese proficiency test prior to sitting the pilot IC test, and the HSK levels the test-takers achieved were HSK 6 (N=5), HSK 5 (N=3) and HSK 3 (N=1). Since the HSK test proper[3] does not have a speaking component in it, the HSK levels test-takers achieved might not be a good indicator of their IC or speaking skills and hence were only used as an indicator of their general proficiency. I aimed to have a range of proficiency levels represented in the 11 pilot test-takers.

In terms of other demographic variables, I tried to ensure the test-takers were as heterogenous as possible within the small sample size, varying in terms of age groups, length of residence, work experience and study experience in the target community, as shown from Table 21 to Table 24. It was hoped that by purposefully selecting a group of L2 test-takers that were diverse in their backgrounds the IC construct and rating scale to be developed could offer more applicability to a wider range of potential test-takers. The fact that some of the pilot test-takers never lived, studied, or worked in China did not imply they were unable to undertake the tasks in the IC test as the task scenarios are of a general, every-life nature without presupposition of specialized knowledge. Even though an L2 test-taker might not have been in a Chinese workplace setting, it is not unreasonable to expect them to have a general schematic understanding of how workplace communication unfolds either in Chinese or in their L1, from either lived or vicarious experiences (e.g., from talking to their parents or watching TV series depicting workplace communication). This is no different from major English language tests assessing test-takers' ability to handle academic discourse

3 There is a separate HSK speaking test called HSKK (Hanyu Shuiping Kouyu Kaoshi), which is not as popular among test-takers as HSK. None of the 11 pilot test-takers took HSKK.

in English when the students have never set foot in an English-medium university. This point is elaborated in detail in Section 4.2.5.

Table 21 Pilot L2-Chinese test-takers' age groups

Age group	N	%
41–45	2	18.18 %
35–40	1	9.09 %
30–35	1	9.09 %
26–30	5	45.45 %
21–25	2	18.18 %

Table 22 Pilot L2-Chinese test-takers' residence length

Length of residence in China	N	%
More than 10 years	3	27.27 %
3–5 years	1	9.09 %
1–3 years	3	27.27 %
Six months to one year	1	9.09 %
Below six months	1	9.09 %
Never	2	18.18 %

Table 23 Pilot L2-Chinese test-takers' work experience

Work experience in China	N	%
More than three years	3	27.27 %
One year to three years	0	0 %
Less than one year	3	27.27 %
Never	5	45.45 %

Table 24 Pilot L2-Chinese test-takers' study experience

Study experience in China	N	%
Studied in China	7	63.63 %
Never studied in China	4	36.36 %

As to the 11 L1-Chinese test-takers in the pilot group, they were all L1 speakers of Chinese who grew up in China, completed tertiary education in China and had a minimum of three years of work experience in China. The inclusion of the L1 group was to ascertain to what extent IC was influenced by test-takers' linguistic competence (LC). It was also hypothesized that the L1-speaker status and L1-speaker LC did not necessarily guarantee high IC (see Sections 2.3 and 2.4 for a discussion on this topic) so L1-speakers were included to test this hypothesis.

5.1.1.2 Pilot test raters

Three L1-Chinese speakers were recruited as raters to offer an intuitive assessment of the test performance of the 22 pilot test test-takers. The raters were in the age range of 45–55, grew up and lived in China and had rich life, study, and work experiences in China. They were members of the Chinese society and possessed implicit knowledge of how interaction should be conducted in a manner that would be considered appropriate in their community. The rationale for rater selection was in line with Hymes's argument that everyday members of a community have the ability to assess the conduct of others based on their lived knowledge of the interactional norms in the community (Hymes, 1972), which is discussed in detail in Section 2.6.2.

5.1.1.3 Everyday-life domain experts

36 everyday-life DEs residing in China were recruited as informants to generate the IC test construct and the rating scale based on their perception of the 22 pilot test-takers' performance data. I selected four DE informants for each one of the nine test tasks, totalling 36 DEs. All DEs had no training or experience in language teaching, assessment or CA, making them CA-naïve raters and linguistic laypersons as defined in Sato and McNamara (2019). The absence of metalinguistic knowledge, however, does not indicate an inability to assess everyday interactional conduct as the DEs were everyday members of the Chinese society and had implicit knowledge of the interactional norms in their community, similar to the argument made in the selection of pilot test raters in the previous section.

Although both the pilot test raters and DEs were L1-Chinese linguistic laypersons, the selection criteria for the 36 DEs were stricter than the ones for the three pilot test raters. For the pilot test raters, I was interested in gathering L1-Chinese naïve raters' intuitive assessment of pilot test-taker performance to ascertain the consistency of L1 speakers' judgement and the functioning of the

test. For this purpose, the pilot test raters only needed to have general experience in and familiarity with the test scenarios. The criteria for the recruitment of the 36 DEs were more restrictive because the DEs were expected to provide expert insight on their respective test items. Therefore each DE's background and experience needed to match the item they were assigned with. To achieve this goal, a referral sampling technique was employed in the recruitment of DEs to identify DEs who were reported by their L1-Chinese peers to be highly experienced in the handling of their respective task scenarios. During initial screenings, all 36 DEs reported that they were in the age range of 25–40, were tertiary educated and had only resided in China. A detailed explanation of how DEs' backgrounds match the task scenarios can be found in Table 25. Details regarding the item content are in Appendix III.

Table 25 DEs' profiles

Item No/ DE ID	**Item scenario descriptions / DE profiles**
Item 1	Offering negative feedback to a junior student's assignment
DE1–4	Postgraduate students reported by their classmates to be highly skilled in mentoring junior students
Item 2	Returning online purchase to a shop owner who is the mother of one's friend
DE5–8	Experienced online shoppers reported to be competent in maintaining positive relationships with shop owners who were their family members or friends
Item 3	Cancelling one's attendance at a team bonding activity organized by one's colleague
DE9–12	Chinese company employees reported by colleagues to be experienced at handling interpersonal relationships in the workplace
Item 4	Informing one's neighbour that one forgot to send off their birthday parcel
DE13–16	Apartment residents reported by their neighbours to be skilled in managing neighbourly relationships
Item 5	Criticizing a colleague one manages when the colleague is related to one's manager
DE 17–20	Chinese company employees reported to have demonstrated team management experiences that involved delicate interpersonal situations
Item 6	Refusing one's university lecturer's invitation to work on the lecturer's project
DE 21–24	University students reported to be skilled at handling relationships with their teachers

Table 25 Continued

Item No/ DE ID	Item scenario descriptions / DE profiles
Item 7	Criticizing one's friend's son who is living in one's apartment
DE 25–28	Apartment residents reported to have maintained positive relationships with friends or relatives whom the residents hosted
Item 8	Disagreeing with one's classmate when working together on a group project
DE 29–32	University students reported to have managed group work very successfully
Item 9	Complaining to one's manager about perceived unfairness from the manager
DE 33–36	Company employees reported to have very positive relationships with their managers

5.1.2 Instruments

The main research instrument employed in the rating material development study is the IC test developed in Chapter 4. A summary of the test can be found in Section 4.2.5 and the test specifications are presented in Table 12 in Section 4.2.2. Pilot raters used a three-point holistic rating scale to rate 22 pilot test-taker performances on all nine items. Details of this process are provided in Section 5.1.3.1.

As to DEs' indigenous criteria, the stimuli used to elicit their indigenous criteria were the same pilot test test-takers' performances on the IC test. Similar to the methodology in Sato & McNamara (2019), DEs were also provided with an instruction sheet on how to group performances based on their indigenous criteria and explicate the rationale behind their decisions. More information on this can be found in Section 5.1.3.2. The inclusion of L1-speaker performance in the stimuli was different from the practices adopted in previous IC assessment studies (Ikeda, 2017; Youn, 2013) as IC assessment research so far focuses on using L2 speakers' performances to develop IC rating scales. The rationale for the inclusion of L1 speakers is explained in Section 5.1.1.1.

5.1.3 Procedures and data analysis

This section, Section 5.1.3, details the steps involved in developing the test construct and rater training materials. The data analyses undertaken in this process are also discussed.

5.1.3.1 Pilot testing

The purpose of pilot testing is to first ascertain if the test can function as intended in the setting for which the test is designed, which means test-takers

can understand the instructions and complete the test tasks in a real testing setting. The feasibility of CMC as a testing platform can also be investigated. Second, the test-taker performance generated by the 22 pilot test test-takers can serve as stimuli materials for DEs to comment on and explicate their indigenous criteria. Third, the pilot test rating of the pilot test performances by the three pilot test raters can generate information about whether the test is assessing a latent unidimensional construct of IC. Lastly, the pilot test test-taker performances can be utilized to create the rating scale and performance exemplars. Performance exemplars are particularly helpful for rater training and can be used as rating aids (see the discussion in Section 2.6.3).

I first administered the IC test to the 22 pilot test test-takers, all of whom completed the nine tasks in the test, totalling 198 test-taker performances. The three pilot test raters rated all 198 performances in a fully crossed design. Their rating was holistic and intuitive in that they were asked to only rate on a three-point scale on the task-taker's interaction: three points for *successful*, two for *average* and one for *unsuccessful*. Raters were instructed to rate based on their intuitive understanding of interactional success, focusing on overall interactional conduct and appropriateness. Pilot test rating results were analysed through Many-Facet Rasch Measurement to ascertain if the three raters had a stable and reliable estimate of interactional success in test-taker performances and whether IC, an unspecified latent test construct at this stage, was unidimensional. If the answer turned out to be affirmative, it would imply the three pilot test raters had a consistent understanding of what IC was. Based on this, it would be reasonable to hypothesize that everyday-life members in a speech community would converge in their judgement of interactional success, which provided a form of empirical basis for the elicitation of everyday-life DEs' criteria of IC at a theoretical level.

5.1.3.2 Eliciting DEs' indigenous IC criteria

Having pilot-tested the test and collected test-taker performances from 22 pilot test test-takers, I sent the test-taker performances to 36 DEs to elicit their indigenous criteria. Each DE was responsible for only one task and needed to only listen to the 22 test-taker performances on that particular task, as shown in Table 25. For example, DE25, DE26, DE27 and DE28 were assigned the 'criticizing your best friend's son' task in item 7 so they only listened to the 22 pilot test-takers' performances on that task, not the other eight tasks the test-takers completed. The advantage of having DEs focus on a particular task is that it reduced the workload for DEs as they only needed to listen to the

performances for one task. At the same time, the DEs could have the capacity to listen to all performances from the 22 pilot test test-takers, which covered a wide range of test-taker profiles and abilities. This approach also ensured the resultant IC test construct represented the views of a larger cohort of DEs, compared to if fewer DEs were recruited. The instructions for DEs were that they should listen to all 22 test-taker performances in the one task they were responsible for and sort the 22 performances into three groups: *successful interaction*, *average interaction,* and *unsuccessful interaction*. There was no requirement for the number of performances in each group and DEs had the option of not putting a performance in any of the groups for any reasons they might have.

After sorting the performances into groups, DEs were required to write brief comments on each of the 22 performances to justify their judgement for sorting. They were encouraged to provide details on which snippet of the interaction they found well managed or poorly managed. Such comments could look like 'at 1 minute 23 seconds the test-taker handled the interaction very well' or 'this line or that expression the test-taker used was particularly conducive to successful interaction'.

Once DEs finished evaluating test-takers' performance individually, I collated their written comments and organized focus group interviews in a paired format. Since there were 36 DEs in total, I conducted 18 paired interviews with each interview consisting of two DEs who were responsible for the same task, which means DE1 and DE2 for task one had a paired interview with the researcher, DE3 and DE4 for task one had an interview, DE5 and DE6 for task two had another interview, etc. The reasoning for conducting the interviews in this manner is that all the DEs were residing in China while I was in Australia, making it logistically challenging to organize bigger focus groups since I was not in China to coordinate. It was easier to find a time that suits two DEs at a time for the interviews.

The interviews with DEs were conducted on video-chat platforms. During the interviews, I played a selection of the 22 performances, which included performances from each of the three performance groups (successful/average/unsuccessful). I purposely selected performances that the two DEs agreed in their judgement (e.g., both DEs considered the performance to be average) and performances on which the two DEs had contradicting views. Performances that DEs found difficult to sort into groups were also selected. I employed stimulated recall by stopping at the snippets in the test-taker performance where DEs made notes in their written comments and allowed DEs to freely discuss the rationale behind their comments. The purpose of the focus group interviews was not to reconcile differences in DEs' perceptions of IC but to allow DEs to

explicate their tacit, implicit members' knowledge of what constituted (un) successful interaction. Adopting the techniques of H-S interviews discussed in Section 4.1.2.1 in Chapter 4, I mediated the discussion and prompted each pair of the DEs to make explicit the concepts they mentioned by asking seemingly self-evident questions such as 'what did you mean by good relationships here'. Though these questions were at times perceived as challenging to answer by DEs, the H-S interview questions helped me and the DEs to elucidate DEs' tacit knowledge and indigenous criteria of IC.

The 18 interview sessions were audio recorded, orthographically transcribed and stored in 18 separate documents, totalling 229,806 words. DEs' written comments prior to the interviews were organized by the test tasks they were responsible for, generating nine separate documents corresponding to the nine tasks, totalling 47,162 words. Combining the interview transcript documents with the written comment documents, I compiled 27 documents in total, which were used as materials for the generation of DEs' indigenous criteria of IC.

5.1.3.3 Developing a DEs' indigenous IC criteria rating scale

After collating DEs' interview transcripts and written comments, similar to the thematic analysis used in the TBNA in Chapter 4, I conducted a thematic analysis (May et al., 2020; Pill, 2016) on both sources of data in NVivo. Extracts of data were first coded and then collapsed into categories that could serve as potential rating categories in the rating scale. A coding scheme of the categories emerged and was revised iteratively as coding progressed. A second coder coded one of the DE interviews using the coding scheme and the percentage of exact agreement with my coding was 82.31 %, which displayed high intercoder reliability.

Having identified DE categories that could potentially serve as rating categories, I conducted two more rounds of coding and thematic analyses within the categories to identify different levels of performance that could translate to steps in the rating scale and to flesh out the descriptors in the rating categories. As DEs grouped test-taker performances into *successful interaction*, *average interaction,* and *unsuccessful interaction*, this information was also used to inform the number of steps in the rating scale. More details of these two thematic analyses can be found in Sections 5.2.3.2 and 5.2.3.3 in the results and discussion section of study two.

After three rounds of thematic analyses, an analytic indigenous IC rating scale emerged which has different rating categories and different steps indexing differing levels of performances, populated by descriptors derived from DEs' comments.

5.1.3.4 Theoretically expanding the IC rating scale

As Section 2.6.2 in the literature review discusses, a rating scale based on raters' perspective, or more specifically in this book, DEs' indigenous criteria, can offer precision in the measurement of test-taker performance based on criteria localized to the particular test tasks and scenarios. The drawback of an indigenous rating scale is that it lacks theoretical underpinnings and generalizability to other assessment contexts. To combat the a-theoreticalness of indigenous scales (Brindley, 1998; Shohamy, 1996), I undertook an additional step to theorize the indigenous rating scale by connecting the descriptors in the DEs' indigenous IC scale with concepts and theories in sociology and philosophy. This process was rooted in the practice where DEs' comments were validated empirically using Sequential-Categorial Analysis (combining CA and MCA) on test-taker performance data. Details of this process are provided in Sections 5.2.3 and 5.2.4. The resultant theorized IC scale, presented in Section 5.2.5, has a stronger theoretical foundation and can be utilized in a wider range of assessment contexts to suit different speaking test tasks and scenarios.

5.2 Results and initial discussion of study two

Section 5.2 presents the findings of study two. Section 5.2.1 first reports the Rasch quantitative results of the pilot test rating, which establishes a preliminary argument for the soundness and unidimensionality of the IC test. A unidimensional IC construct, even unspecified and based on pilot raters' intuition, is crucial to the specification of the construct by DEs. Section 5.2.2 details the results of the thematic analysis of domain experts' indigenous criteria, which were then collapsed into an indigenous IC rating scale in Section 5.2.3. Section 5.2.4 qualitatively validates DEs' indigenous rating categories, drawing on DEs' comments and Sequential-Categorial Analysis of test-taker discourse. This process also generated CA and MCA-informed test-taker response exemplars, which were used for rater training in study three in Chapter 6. Finally, Section 5.2.5 theorizes DEs' indigenous rating scale in Section 5.2.3 into a CA and MCA-informed theorized scale, based on the analyses in Section 5.2.4, while Section 5.2.6 presents a unified model of the IC construct – a highlight of this book – that captures previously under-theorized dimensions of interaction. Parts of the results were presented at regional and international peer-reviewed conferences (Dai, 2019a, 2019b, 2019c).

5.2.1 Pilot test findings

As Section 5.1.3.1 explains, the three pilot test raters gave a score to each of the nine performances from the 22 pilot test test-takers in a fully-crossed rating

design. This holistic rating scheme assigned 3 for a *successful* interaction, 2 for an *average* one and 1 for an *unsuccessful* interaction. This generated rating data that were analysable by Rasch to examine the functioning of the items and raters' ratings. Figure 13 presents the Rasch Wright map with all the facets of the pilot rating results.

Rasch analysis of the three raters' ratings on the 22 test-takers showed that test-takers were spread over a wide range of ability levels. Fair scores, which are scores corrected for rater severity and item difficulty through Rasch, are presented in Table 26. The scores in Table 26 are test-takers' average fair scores across all nine items. The table also includes test-taker IDs and whether they were in the L1-Chinese speaker group or the L2-Chinese speaker group. Looking at Table 26, though L1-Chinese speakers dominated the higher end of the scale, many L1-Chinese speakers did not outperform L2-Chinese speakers in the middle range of the scale and one L1-Chinese speaker (ID 13) even scored below all 11 L2-Chinese speakers. Within the middle range, it is telling to note that test-taker 12 was a beginner-level L2-Chinese speaker, overtaking not only five L1-Chinese speakers but also eight L2-Chinese speakers of much higher Chinese proficiency. Such findings offer emergent quantitative evidence suggesting that the IC captured by the three pilot test raters' holistic judgement of interactional success is different from LC, which is traditionally benchmarked against L1 speaker competence and measured in proficiency testing such as HSK. Although the six highest scoring test-takers were all L1 speakers, L1 speakers were not invariably better-performing than L2 speakers. If IC is measuring exactly what is measured by LC, we would be seeing L1 speakers score higher than L2 speakers in all cases, which is not true from this pilot. The pilot study suggests that when raters were asked to rate successful interaction, they were not solely considering the criteria of LC. What IC criteria they drew on was unclear at the stage of pilot testing. The specification of the IC construct that underlies everyday members' criteria was achieved through the elicitation of DEs' indigenous criteria in subsequent steps.

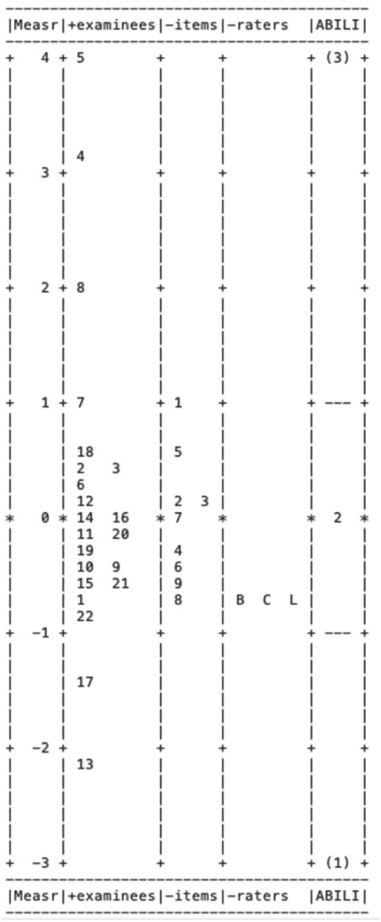


Figure 13 Wright map from the pilot testing dataset

Table 26 Pilot test test-takers' fair scores

ID	Group	Score
5	L1	2.99
4	L1	2.97
8	L1	2.90
7	L1	2.73
18	L1	2.59
2	L1	2.55
3	L2	2.55
6	L2	2.47
12	L2	2.43
14	L2	2.36
16	L1	2.36
11	L2	2.28
20	L1	2.28
19	L2	2.20
9	L2	2.16
10	L1	2.16
15	L2	2.08
21	L2	2.04
1	L2	1.95
22	L1	1.91
17	L2	1.60
13	L1	1.34

Rasch analysis also generated findings regarding the performance of the nine items in the test and indexes of rater reliability and test reliability. The Rasch test-taker reliability (similar to test reliability such as Cronbach's alpha in classical testing theory) was 0.85, which is high enough for a small-scale pilot test. All nine items were within the infit range of 0.77 to 1.31, suggesting that the items were measuring a unidimensional construct, which is IC in this test. The three raters demonstrated acceptable intra-rater reliability, evidenced by their infit values, which were 0.98, 1.01 and 1.03. When the infit values are neither misfitting nor overfitting, they suggest that raters can rate consistently in their ratings according to the expectations of the Rasch model. The fair averages of the three raters were also the same (2.37 for all three raters), suggesting consistency across raters and high inter-rater reliability. Finally, the rating scale functioning showed

that the raters utilized all three steps (*unsuccessful*, *average,* and *successful*) in the scale, with 117 times of *unsuccessful* awarded to test-taker performances, 180 times of *average* awarded, and 270 times of *successful* awarded. The measures for the three steps were -0.30, 0.39 and 1.30 in logits. The distances between the measures revealed that the ability range between *unsuccessful* and *average* (0.69) was smaller than the range between *average* and *successful* (0.91). This indicates that more steps in the rating scale are needed to capture the differing levels from *unsuccessful* to *successful* interaction.

In summary, although the pilot test Rasch results are based on a very small dataset, an unspecified rating scale (unsuccessful, average, and successful), and raters' intuition without any rater training, looking at the Rasch results, there are reasons to believe that the test was functioning as expected and could spread test-takers across different ability levels. The three raters reliably assessed what they considered to be successful interaction, although at this stage what criteria they used for successful interaction are unknown. The high inter- and intra-rater reliability indexes in the pilot study suggest that linguistically untrained raters had the ability to assess speakers' IC in a very consistent manner. This shows that it is possible to elicit everyday-life members' understanding of interactional success to unpack the test construct. Based on this insight, I proceeded to the next step of study two, which was to elicit everyday-life DEs' indigenous criteria of IC.

5.2.2 Domain experts' indigenous IC criteria

After conducting the pilot test, I elicited DEs' indigenous criteria based on pilot test takers' performances as explained in Section 5.1.3.2. Similar to the reporting of the findings of the TBNA in Chapter 4, the coding results of DEs' comments are presented in the format of the number of coded extracts, codes and categories. The coded extracts were direct quotes from the written comments and interview transcripts produced by DEs. Codes were assigned to extracts that shared similar content. Categories further abstracted related codes and indexed different aspects of IC in DEs' perception. The naming of codes mostly followed the language used by DEs to avoid undue influence from me, the researcher, while the naming of categories, due to their abstracted nature, used more theorized terms based on my judgement. Table 27 presents the names of the categories, the number of coded extracts in each category, the percentage of extracts in each category out of the total number of coded extracts, and the number of files from which the extracts in each category were derived. There are in total 27 files consisting of 18 interview transcript files (corresponding to the 18 interviews) and 9 written comment files (corresponding to the nine items),

as Section 5.1.3.2 explains. The reason for providing information on the number of files from which the categories were developed is to show the coverage of a particular category. The more files a category covers, the more applicable and generalizable that particular category is. In Table 27 the categories are ranked based on the number of coded extracts in them.

In the following sections, a brief summary of what each category entails is presented. Due to space restrictions, details of DEs' comments and the relevant test-taker performances are not presented. They are instead selectively represented in Section 5.2.4, where the focus is on combining DE comments with Sequential-Categorial Analysis of test-taker performance to empirically validate the rating categories.

Table 27 Indigenous categories coverage

Categories	Coded extracts	Percentage	Number of files
Conflict management	877	17.59 %	27
Solidarity promotion	805	16.15 %	27
Reasoning skills	729	14.62 %	27
Personal qualities	729	14.62 %	27
Social relations	724	14.52 %	26
Linguistic choices	276	5.54 %	24
Prosodic features	274	5.50 %	24
The structure of talk	225	4.51 %	26
Strategies	129	2.59 %	24
Miscellaneous	103	2.07 %	20
Efficiency	70	1.40 %	18
Cultural norms	45	0.90 %	14
Total	4,986	100 %	27

5.2.2.1 Conflict management

Since the test tasks elicit test-takers' ability to handle disaffiliative situations, it is understandable that DEs oriented most frequently to features in test-taker performance that indexed their ability to manage conflict/disaffiliation. Before the content in the *conflict management* category is explicated, a brief note on how the content will be presented is in order. For each category, I detail the most mentioned codes in that category, with the names of the codes italicized. A bracket follows the code with the first number in the bracket indicating the number of coded extracts using that code, and the second number showing

the number of files a particular code covers. A code with a bigger first number means that the code has a larger number of extracts coded with this code, while a code with a bigger second number indicates that the code encompasses a larger number of files out of the 27 files for the thematic analysis.

For *conflict management*, the codes centre on test-takers' ability to use *indirect language and action* (221/23), to use language that is *moderate* (96/18), and to *circumvent direct and forceful actions* (86/18). Social actions that DEs found particularly problematic are actions that are *threatening* (61/11), *questioning and rebutting* other people (43/10), *complaining* about others (40/3), *excessively criticizing* others (16/6) and *escalating the problem* (36/6). Though not every code in this category is detailed here, it should be clear that the category *conflict management* depicts test-takers' competence in using indirect language and action to mediate disaffiliative situations. DEs favoured behaviours that index the ability to do so and sanctioned approaches that are overly disaffiliative.

5.2.2.2 Solidarity promotion

Contrary to conflict management, the category *solidarity promotion* describes test-takers' ability to use social actions that affiliate with other people so as to promote social solidarity. Such actions include *enquiring about other people affectionately* (101/14), *negotiating and discussing to reach an agreement* (101/17), *demonstrating understanding* (52/11), *approving and affirming other people's ideas* (49/12) and *offering encouragement and good wishes* (40/7). Empathy was also highlighted by DEs as they oriented to the ability to *look after the other person's feelings* (67/18) and *put oneself in other people's shoes* (60/19). Solidarity promotion differs from conflict management in that the latter focuses on the ability to circumvent or downgrade disaffiliation while the former highlights the ability to actively promote solidarity and affiliation.

5.2.2.3 Reasoning skills

Since all test tasks are based on real-life interactional situations, the tasks require a high degree of problem-solving skills from the test-takers as they navigate the interactional challenges. DEs emphasized this ability, which was conceptualized as *reasoning skills* in this category. Reasoning is the ability to be *solution focused* in one's approach (191/23) and to be able to supply *reasonable and sensible solutions* when there were interactional troubles (156/22). When a test-taker needs to provide reasons for their actions, such as when explaining why one did not send a parcel for their neighbour as promised (item 4 in Appendix III) or cannot attend a team-bonding activity they signed up for

(item 5 in Appendix III), DEs expected their reasons to be *reasonable* (145/20) and *sufficient* (75/15).

5.2.2.4 Personal qualities

Personal qualities is a category that at first sight can seem unrelated to language skills or IC. However, when interaction takes place, we unavoidably make inferences about other interactants' characters, which is the concept of morality discussed in CA literature (see Section 2.4.4 and more specifically, Bergmann, 1998 and Turowetz & Maynard, 2010). This is an aspect of interaction that was frequently commented on by DEs. Some of the positive personal qualities valued by DEs are *being sincere* (109/19), *being honest* (80/12), *being respectful* (50/17), *being collaborative and team-spirited* (49/8), *being flexible* (47/14) and *being proactive* (47/6). Some of the negative personal qualities noted by DEs are *being self-centred* (19/5), *being vengeful* (19/4), *harbouring malicious motives* (14/5) and *imputing negative intentions in others* (10/5)

5.2.2.5 Social relations

As discussed in Section 2.4.6 in the literature review in Chapter 2, no language use is unbounded to social roles as interactants are constantly doing *being a student*, *being a friend* or *being an employee* in different interactional settings. The wide range of interactional settings (N=9) in this IC test elicits test-takers' ability to enact various social roles and design their language to the different social roles their interlocutors have. The manners in which DEs perceived this ability are in their comments about *choosing language that is suitable to one's roles* (198/21), *choosing language that is suitable to the roles of other people* (195/24) and *showing a good understanding of the seniority and social ranking between oneself and other people* (195/24). This shows that DEs not only expected test-takers to appreciate their own social roles in the test task scenarios but also to understand how to tailor their talk to their interlocutors' social roles, with an awareness of the interpersonal social hierarchy and relations.

Another aspect of *social relations* that DEs oriented to is the ability to match one's action with the social roles in the task scenarios. DEs expected test-takers to *fulfil the responsibilities and obligations in one's roles* (65/10), to *make requests that are befitting one's roles* (44/11) and to *understand one's rights* (34/7).

5.2.2.6 Linguistic choices

The three categories, *linguistic choices*, *prosodic features,* and *the structure of talk* from Sections 5.2.2.6 to 5.2.2.8, bear a stronger resemblance to traditional LC

assessment criteria. The difference between these three categories and more psycholinguistically-oriented LC categories such as *lexical range*, *pronunciation* and *cohesion* is that DEs did not treat them as the goals of interaction. In other words, neither linguistic choices, prosodic features, or the structure of talk was perceived as markers worthy of commenting in themselves. Instead, they were discussed by DEs in relation to the interactional import they had. This propensity has been documented in previous studies that used linguistic laypersons as informants for test constructs (Sato & McNamara, 2019). This also mirrors the results in the TBNA in Chapter 4 where the TBNA participants predominantly commented on non-linguistic/non-LC factors when the topic under discussion was challenges in interaction (see Section 4.2.1).

For linguistic choices, which in this context is related to how social actions are organized lexically, DEs frequently commented on how the *choice of word* (98/18), the *use of hedging devices* (45/11) and the *use of sentence-final particles* (26/11) and the *use of intensifiers* (21/13) could impact on the success or lack of success of interaction. It should be noted that DEs did not have the metalinguistic knowledge to pinpoint and describe such linguistic features. They commented on the effect of the use of this or that word and these effects were thematically coded by the researchers with labels such as *stance*, *hedges* and *modal verbs*, depending on the functions the words served in the interaction.

5.2.2.7 Prosodic features

Prosodic features is another more linguistically oriented category and DEs primarily focused on *intonation* (226/24). Similarly, due to DEs' lack of training in linguistics and their impressionistic listening of performance, they did not have the insight from metalinguistic knowledge or more refined phonetic or prosodic analyses of test-taker talk. Therefore, it was difficult to further classify DEs' every single use of the term *intonation*, which could encompass a range of prosodic and phonetic features. Other more explicitly mentioned *prosodic features* include *pitch* (34/14), *speech rate* (3/3) and *stress* (3/2). It is clear that *prosodic features* were factored in DEs' judgement of interactional success.

5.2.2.8 The structure of talk

The category structure of talk is similar to *coherence* and *task fulfilment* in traditional speaking assessment criteria, though differences exist in DEs' focus in this study. Features that were most commented on included *fulfilling the main interactional task and expressing one's ideas fully* (52/6), *comprehensiveness in one's*

narration and response (40/17), *the provision of prefaces* (33/12), *being structural and organized in one's talk* (29/12) and *expansion in one's narration* (28/12).

5.2.2.9 Strategies, cultural norms, and miscellaneous

Due to the relatively fewer codes, the last three categories are discussed together in this section. The *strategies* category encompasses specific interactional strategies that DEs noted in test-takers' performances. Such strategies can be highly context-specific such as *soliciting information before action* (18/5) and *swallowing one's pride to achieve goals* (6/4), which were only applicable to some of the nine test tasks but not others. These strategies serve the purpose of mitigating disaffiliation and can be incorporated into the *conflict management* category in more general terms. The codes in the *cultural norms* category are similarly specific to only certain items or interactional settings such as *avoiding excessive discussion of money* (6/2) and *maintaining boundaries between the opposite sex* (3/1). The *miscellaneous* category is mostly comprised of mentions of general interactional/pragmatic terms such as *being polite and appropriate* (94/19). Other codes in this category include *showing high intelligence* (2/1) and *language is modern* (1/1), which, due to their sparse mentions, did not warrant a focus in DEs' indigenous criteria.

In summary, considering the much smaller numbers of codes in the last three categories and the much less emphasis placed on them in DEs' comments, it is appropriate to either incorporate them into existing categories or exclude them in the selection of rating categories for a rating scale.

5.2.3 An indigenous IC rating scale

5.2.3.1 Collapsing indigenous criteria into five rating categories

Having identified 12 categories that DEs oriented to in relation to IC in Section 5.2.2, this section focuses on how the top-mentioned nine categories were mapped onto an IC rating scale that was based on DEs' indigenous criteria. A nine-category rating scale clearly is cumbersome, impractical and cognitively overloading for raters. Further abstraction and simplification are needed to narrow down the IC categories to create an operational rating scale.

As discussed in Sections 5.2.2.6 and 5.2.2.7, the two categories, *linguistic choices* and *prosodic features*, represent the linguistic and paralinguistic devices test-takers use to realize social actions. Since the management of conflict (Section 5.2.2.1) and the promotion of solidarity (Section 5.2.2.2) are more concerned with language use compared to other categories, it is reasonable to subsume categories *linguistic choices* and *prosodic features* under categories *conflict*

management and *solidarity promotion*. *The structure of talk*, discussed in Section 5.2.2.8, indexes test-takers' ability to organize talk in a cohesive, well-reasoned manner, which shares with the *reasoning skills* category in their requirement of logic and reasoning competence in test-takers. With *the structure of talk* category incorporated in *reasoning skills*, the *strategies* category subsumed under *conflict management*, and the *cultural norms* and *miscellaneous* categories removed (see Section 5.2.2.9), we now have a five-category conceptualization of IC based on DEs' indigenous criteria, as Table 28 illustrates.

Table 28 Indigenous rating scale five rating categories

Categories	Number of files the categories were found
Conflict management	27
Reasoning skills	27
Solidarity promotion	27
Personal qualities	27
Social relations	26

The five categories in total account for 93.03 % of the entirety of coded extracts, offering comprehensive coverage of DEs' conceptualization of the test construct with minimal 'construct shrinkage' (Knoch et al., 2020, p. 1). The fact that the five categories are found in nearly all the DE interviews and written notes documents (N=27, see Section 5.1.3.2) also indicates that there is strong generalizability of the five rating categories as each of them sufficiently covers all the nine test tasks. In other words, the five rating categories represent IC markers that are relevant to all the items in this test, whether the items are in the everyday interactional, academic or workplace contexts. This finding is encouraging as existing IC assessment criteria tend to utilize IC markers that are context- or task-specific such as *topic development* for pair discussion tasks. Though localized, context-fitting IC markers provide strong precision in measuring task-relevant features, such markers cannot be easily extrapolated to other assessment tasks, reducing the practicality of IC assessment. Guided by DEs' judgement and the inductive thematic coding process, the five IC rating categories this study arrived at have greater potential to be applied to a wide range of assessment settings and task types that are similar to the ones in the IC test in this book. The large number of DEs (N=36) recruited in this study also shows that these five rating categories represent IC criteria that are valued by the general everyday members of society, reflecting the prevailing opinions of interactional abilities.

5.2.3.2 Identifying steps in the rating categories

Having arrived at a five-category abstraction of DEs' indigenous criteria,for the next step in the rating scale design I needed to identify steps in each category. Although the previous NVivo thematic analysis results explicate the distribution of codes within each category and make clear which codes were more frequently commented on and which ones were less salient to DEs (see the number of coded extracts and number of files for the codes in each category in Section 5.2.2), this thematic analysis does not differentiate the differing qualities of performances in the categories, such as *strong*, *average* and *not-so-good* performances within the category *conflict management*. In other words, the thematic analysis in Section 5.2.2 can inform us of what *conflict management* entails, but not what a *good* conflict management is vis-à-vis a *not-so-good* one. This type of information is needed for the creation of scale steps as the steps in a scale embody levels of differences in performance.

In order to address this issue, I went back to the fair scores from the pilot rating (see Section 5.2.1) and the results from DEs' sorting exercise (see Section 5.1.3.2). The fair scores produced by pilot raters provided an estimate of the performance range from the 22 pilot test-takers. As Figure 13 illustrates, the 22 test-takers spread across more than 6 logits, indicating that more steps were needed for more precise differentiation of ability levels. Apart from quantitative evidence in support of more steps in the rating scale from pilot rating, DEs' comments provided qualitative evidence pointing in the same direction. Before commenting on test-takers' performances, DEs were required to put their respective 22 pilot-test performances in three groups, *successful*, *average* and *unsuccessful*. Therefore, there were initial clues in DEs' comments to assist me with the differentiation of different levels of performance. Having re-read DEs comments, I noticed DEs made frequent observations that it was challenging to pigeonhole performances into the prescribed three groups (*successful*, *average,* and *unsuccessful*) as there were performances that were in-between. Some performances were better than *average* but paled in comparison to *extremely successful* ones. There were other performances that were decidedly below *average* but were of different degrees of *unsuccessful*. This suggested that the three steps (*unsuccessful*, *average,* and *successful*) used in the sorting exercise could not directly translate to three steps in the rating scale, as they were unable to capture the finer granularity of the wide range of performances. Informed by the pilot Rasch analysis and DEs' intuitive comments, I decided to employ a five-step solution to differentiate the quality of test-taker performance: *exemplary*, *good*, *average*, *development needed,* and *intervention needed*. Instead of having only one step (*unsuccessful*) under *average*

and only one step (*successful*) above *average*, the new five-step model helped to differentiate the differing levels of *unsuccessful* and *successful* that DEs' noticed. The labels chosen for the five steps were positively phrased to improve test-user engagement.

While probing DEs' differentiation of different levels of performance, I noticed when DEs commented on the quality of a performance, they also noted how a particular level of performance would be received in real-world settings and what type of support and training would be required. This linkage between assessment and real-world implications through the perspective of DEs makes results of this IC test more informative to end-users. Table 29 presents how a performance at a particular step in the rating scale is related to real-world consequences and pedagogical needs.

Table 29 Band levels, real-world correspondences, and pedagogical implications

Step	Definition	Real-world correspondence	Pedagogical implication
5	Exemplary	Performance that is highly praiseworthy. The test-taker can skilfully handle interactional challenges in a wide range of interactional settings. The test-taker will thrive in the target community due to their high IC.	No training needed IC-wise.
4	Good	Performance that the majority of everyday members in the target community would consider to be good. The test-taker can achieve most of their goals and interactional needs to a satisfactory level in the target community.	Sufficient IC, though fine-tuning can be beneficial.
3	Average	Performance that is acceptable and passable. There can be both positive and negative points to the performance. The test-taker can handle most of the interactional tasks in the target community, though not always to a satisfactory level.	An acceptable degree of IC, more teaching and training in IC can be helpful.
2	Development needed	Performance that contains noticeable breaches of interactional norms. The test-taker is likely to offend other interactants.	Insufficient IC; the test-taker is advised to take up training in IC.

(*Continued*)

Table 29 Continued

Step	Definition	Real-world correspondence	Pedagogical implication
1	Intervention needed	Performance that is highly problematic in terms of IC. The test-taker is very likely to encounter major interactional challenges.	The test-takers should undertake substantive and extensive training in IC.

Having identified five distinguishable steps and established their linkage to real-world consequences and pedagogical implications based on DEs' comments, I conducted a second thematic analysis to sort their comments for each rating category into five distinctive levels as identified in Table 29. This process is illustrated in Figure 14.

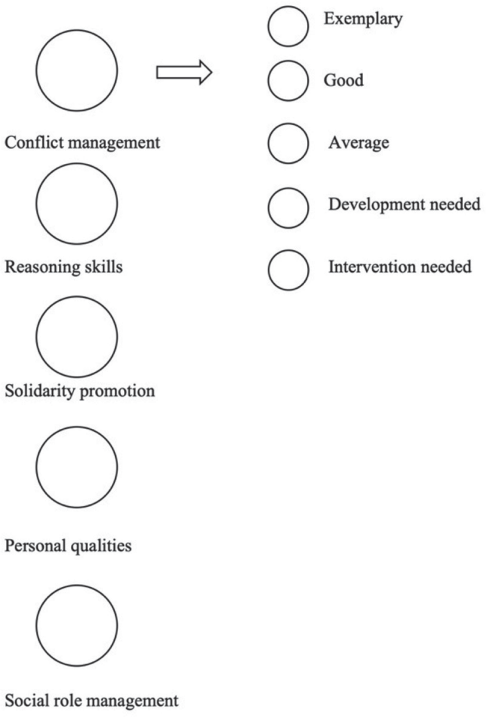


Figure 14 Sorting DEs' comments into five steps

Note. This process was conducted for all five categories though only the one for *conflict management* is represented here.

I first created five separate documents, each of which contained all the coded DE comments for each of the five rating categories in Table 28. For example, the document for the category *conflict management* contained all the codes related to this category from all 27 DE comment documents, 18 of which were from the interviews and nine of which were from the written notes. The connection between the first thematic analysis in Section 5.2.2 and the current second thematic analysis is that the first thematic analysis started with DEs' comments in all 27 comment documents, followed by an inductive process where I arrived at the five broad rating categories and the codes that fell under each of the five rating categories. The second thematic analysis, however, started with the identified five rating categories and sorted all 36 DEs' comments for each one of the five rating categories into the five rating steps. If we take the rating category *conflict management* as an example, I first generated a document that had all the codes related to conflict management from the 27 DE comment documents, which covered all 36 DEs' comments on all nine tasks. I then put this *conflict management* document in Nvivo, created five steps (*exemplary* to *intervention needed*), and sorted all the comments into the five steps. The coding result, therefore, was an NVivo document that differentiated test-takers' handling of disaffiliation by the five steps as specified in Table 29. The same procedure was conducted for the other four rating categories. This inductive, data-driven approach to step creation generated steps that were not based on my *a priori* intuition. Instead, the number of steps and the differences between steps were empirically supported and validated by DEs' judgement as the five steps are what DEs considered necessary and were able to distinguish.

5.2.3.3 Identifying sub rating categories and extracting descriptors

The upshot of the second thematic analysis was five NVivo documents that corresponded to the five rating categories. Each NVivo document contained DEs comments on a specific rating category from all nine tasks with the comments sorted into five steps. The next task in the rating scale design was to write the descriptors for each step. A revisit to the codes and code distribution in the indigenous criteria in Section 5.2.2 revealed that the various codes in each category were not equal in their distribution and therefore importance. The most frequently mentioned codes reported in Section 5.2.2 deserve more prominence in the band descriptors as they were more salient and considered more relevant by DEs. Therefore, the third thematic analysis, which was the extraction of descriptors for each of the five steps in the rating scale, needed to be guided by

the distribution of mentioning of rating-category-relevant features in the first thematic analysis.

The third thematic analysis started with a reread of DE comments in each step in each rating category, produced by the second thematic analysis in Section 5.2.3.2. I then extracted rating-category-relevant descriptors, which were specific to each step in each of the five rating categories. The descriptors were subsequently grouped into sub-categories that typified a rating category. For example, when reading through DE comments for the *exemplary* band in the *conflict management* category, I noticed DEs used descriptors that could be grouped into four sub-categories: *use of face-threatening social actions*, *linguistic control*, *paralinguistic control*, and *mitigating strategies*. I then moved on to the DE comments for the other four bands and revised the definitions of the identified four sub-categories from the top band. The abstraction of sub-categories was also guided by the numbers of coded extracts in the codes reported in Section 5.2.2. For example, the most mentioned code in *conflict management* was *direct and forceful actions* (86/18), so a sub-category focusing on face-threatening actions was considered warranted.

Once the sub-categories for the *conflict management* category were consolidated, I, band by band (from *exemplary* to *intervention needed*), sorted DEs' comments into the four sub-categories: *use of face-threatening social actions*, *linguistic control*, *paralinguistic control*, and *mitigating strategies*. The same process was applied to the other four rating categories. Building on Figure 14, Figure 15 shows how the third thematic analysis was conducted following the second thematic analysis. This methodical, data-driven approach ensured that the resultant sub-categories in each rating category represented the features that were most relevant and noticeable to DEs, which, by extrapolation, would likely be most relevant and noticeable to test raters and users. The descriptors in the sub-categories were also empirically derived from DEs' comments.

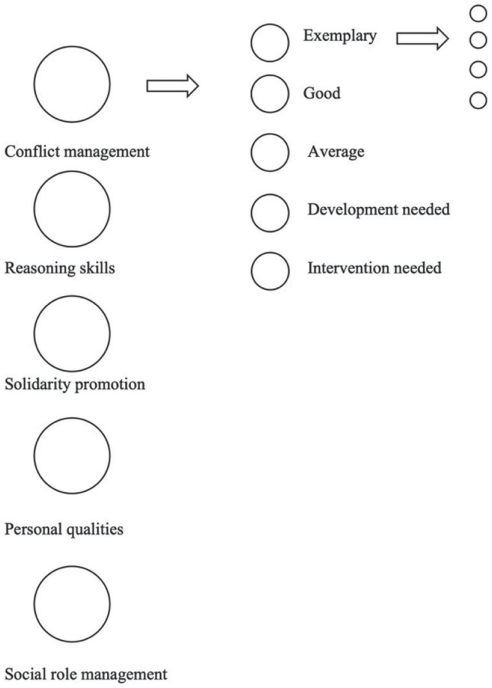


Figure 15 Sorting DEs' comments into sub-categories

Note. This process was conducted for all five bands in each of the five categories, though only the one for the *exemplary* band in the category *conflict management* is illustrated here. The four smallest circles represent the four sub-categories: *use of face-threatening social actions*, *linguistic control*, *paralinguistic control*, and *mitigating strategies*.

The following sections, from Section 5.2.3.4 to Section 5.2.3.8, present the indigenous IC rating scale that has five rating categories, each of which has its sub-categories and descriptors. It should be noted in the write-up of descriptors I aimed to preserve the original language and description from DEs to the fullest possible extent, but the use of some linguistic terminology was necessary to precisely capture DEs' descriptions. The use of CA or MCA terminology was kept to a minimum as CA and MCA were based on an emic understanding of test-taker discourse, which was different from DEs' etic perspective of test-taker performance. After the presentation of the indigenous rating scale, Section 5.2.4 explains how I use test-taker discourse as a common denominator to connect the CA/MCA emic understanding with the DE etic understanding, in an attempt to theorize the indigenous IC scale into a theorized IC scale.

5.2.3.4 Indigenous rating category: Conflict management

Four sub-categories were identified in the rating category *conflict management*, which are *implementation of face-threatening actions (IFA)*, *use of indirect linguistic devices (ILD)*, *use of indirect paralinguistic devices (IPD),* and *use of indirect strategies (UIS)*. Each sub-category has five steps to differentiate different levels of performance. Table 30 presents the *conflict management* rating scale with the abbreviations of the sub-categories listed at the end of the band descriptors.

Table 30 Indigenous category: Conflict management

Indigenous category: Conflict management	
Band/step	**Descriptors**
Band 5: Exemplary	• Face-threatening actions such as criticizing, complaining, questioning, and rebutting are either avoided, downplayed or veiled. Behaviour that is damaging to interpersonal harmony is managed carefully (IFA). • Careful choice of words to mitigate damage to the interlocutor's face. A range of softening devices such as epistemic hedges, modal verbs and sentence-final particles are skilfully employed (ILD). • Excellent control over prosodic features to mitigate damage to the interlocutor's face and convey the message skilfully (IPD). • Employs a range of strategies to enhance indirectness, de-escalate and preserve faces (e.g., hinting, sub-texting, third partying and staying middle-of-the-road) (UIS).
Band 4: Good	• Strong face-threatening actions (e.g., criticizing) are indirectly delivered whereas weak face-threatening actions are delivered in a manner that might appear slightly blunt but is still contextually acceptable (IFA). • Indirect language and softening devices are utilized to convey the message in a non-confrontational and non-accusatory manner. Infelicitous choices such as imperatives might exist (ILD). • Prosody (e.g., pitch and speed) is well managed so the message is delivered in a moderate and composed manner (IPD). • Indirect and strategic in one's approach, though lacking in more elaborate moves (UIS).

Table 30 Continued

Indigenous category: Conflict management	
Band 3: Average	• Face-threatening actions are hearable although social harmony is not disrupted. Inconsistency in one's management of face-threatening actions: some are delivered rather bluntly whereas others are indirectly delivered (IFA). • Choice of words is average, lacking in tact, agility and sophistication. The message needs to be moderated more with softening devices (ILD). • Intonation is acceptable, suggesting that one is in control over their emotions (IPD). • Approach is quite direct and unsophisticated, which can strain the relationship (UIS).
Band 2: Development needed	• Explicit use of face-threatening actions (e.g., negating, criticizing, refusing and refuting) in a manner that damages interpersonal relationships (IFA). • Language is unmitigated, direct and aggressive. Softening devices are underused (ILD). • Intonation is inappropriate, sounding dismissive and aggressive (IPD). • Approach is very forceful and can escalate tension (UIS).
Band 1: Intervention needed	• Strong, overt, and unnecessary display of face-threatening actions that severely disrupt social harmony (IFA). • Poor choice of words delivering an offensive and critical tone. Evidence of bluntness created by linguistic devices (e.g., modals such as 'must', 'shouldn't' or 'can't', imperatives, negation, and intensifiers). Softening devices largely missing (ILD). • Intonation is offensive with little to no softening effect in prosodic features (IPD). • Approach is offensively direct, aggressively persistent, and unnecessarily confrontational. Instead of being calm and collected, the speaker appears hasty, agitable and opinionated (UIS).

5.2.3.5 Indigenous rating category: Solidarity promotion

Four sub-categories emerged from DEs' comments within the rating category *solidarity promotion*. The sub-categories are *being caring and supportive (BCS)*, *use of face-enhancing actions (FEA)*, *demonstrating empathy (DEY)* and *adopting a positive approach (APA)*. Each sub-category and its band descriptors are explained and presented in Table 31 with sub-category abbreviations provided.

Table 31 Indigenous category: Solidarity promotion

Indigenous category: Solidarity promotion	
Band/step	**Descriptors**
Band 5: Exemplary	• Skilfully promotes solidarity by using language and intonation that are caring, supportive, affable, warm, and face-saving for other people (BCS). • Effectively employs face-enhancing social actions such as agreeing, affirming, acknowledging, praising, thanking, trusting, and apologizing to strengthen positivity and social harmony (FEA). • Demonstrates strong empathy and sympathy by putting oneself in others' shoes and showing consideration for their feelings and situation. Can appeal to others' emotions and relate to them on an emotional level (DEY). • Shows keenness in discussing and negotiating with the interlocutor to preserve solidarity (APA).
Band 4: Good	• Demonstrates the ability to use supportive language to show one's care for other people, though it is not always consistent (BCS). • Can utilize commonly-used face-enhancing actions such as agreeing and affirming but may lack in other types of actions such as encouraging and expressing good wishes (FEA). • Demonstrates enough empathy by thinking from others' perspectives. There is evidence of being on the same wavelength though it is not always consistent (DEY). • Sufficient evidence of adopting an approach that focuses on communication, discussion, and negotiation (APA).
Band 3: Average	• Can use some supportive and caring language to promote solidarity but the quality or quantity of such language is often lacking (BCS). • Shows awareness of using face-enhancing actions but such actions are not realized to the desirable extent (FEA). • Shows attempts at being empathetic and understanding but more work is needed to achieve common grounds (DEY). • Leaves room for negotiation and further discussion, though the speaker doesn't actively promote it (APA).

Table 31 Continued

Indigenous category: Solidarity promotion	
Band/step	**Descriptors**
Band 2: Development needed	• Language and intonation lack warmth and support, which can potentially hurt other people's feelings (BCS). • Little attempt at promoting solidarity and the speaker is perceived as not making enough effort in showing support or staying positive (FEA). • Interaction can appear impersonal and unempathetic at times. Other people's feelings and situations are not fully taken into account (DEY). • Focuses on the speaker's perspective, providing little leeway for discussion or negotiation (APA).
Band 1: Intervention needed	• Language is cold, indifferent and face-losing for other people, showing little to no awareness in preserving solidarity (BCS). • Scant use of face-enhancing actions, making others feel that the speaker does not care about their feelings or trust them (FEA). • Causes emotional distress for others by using language that can make them feel worried, scared, hurt, embarrassed or humiliated. Fails to demonstrate attempts at being empathetic or desiring common grounds (DEY). • Supplies little to no room for negotiation or consensus, closing the door of communication (APA).

5.2.3.6 Indigenous rating category: Personal qualities

The third thematic analysis returned four sub-categories within the rating category *personal qualities*, which are *moral character (MCR)*, *demonstrating goodwill (DGW)*, *demonstrating cultural-specific personal qualities (CSQ)*, and *showing respect (SRT)*. Details of the four sub-categories and their band descriptors with sub-category abbreviations are illustrated in Table 32.

Table 32 Indigenous category: Personal qualities

Indigenous category: Personal qualities	
Band/step	**Descriptors**
Band 5: Exemplary	• Possesses strong moral character through the demonstration of virtues such as honesty, integrity, accountability, tolerance, conscientiousness, self-discipline, dependability and dutifulness. Such qualities are very noticeable in the speaker's talk (MCR). • Shows a lot of goodwill by being very sincere, helpful, trustful, positive, and proactive in one's talk and demeanour. The speaker is perceived as a very kind-hearted person (DGW). • The speaker is very humble, modest, flexible, collaborative, and collectivistic in their thinking. It is clear that the speaker focuses on the big picture instead of their own interests (CSQ). • Maintains utmost respect and dignity for oneself and others, even when under stressful situations. Shows thoughtfulness to other people's opinions and situations (SRT).
Band 4: Good	• The speaker shows some positive moral values as listed in the Band 5 descriptors, but the values do not come across as strongly as performances at Band 5 do (MCR). • Can evince sincerity through one's talk and disposition. the speaker is considered in general a good person. The degree of goodwill is less than the one in Band 5 performances (DGW). • There is an adequate amount of team spirit, flexibility, cooperation and collectivism in the speaker's approach. There's no reason to believe the speaker will only prioritize their own interests (CSQ). • Shows respect for other people and their opinions (SRT).
Band 3: Average	• No clear evidence of the possession of strong moral values. Neither is there any display of morally questionable behaviours (MCR). • More needs to be said or done to convey goodwill. There are signs of sincerity, but the speaker is not being helpful, positive, enthusiastic, or proactive enough (DGW). • There is a slight lack of team spirit and cooperation as required in the scenario. The speaker is not being as collaborative and flexible as one would hope for (CSQ). • An acceptable amount of respect is shown but the speaker can appear slightly too casual (SRT).

Table 32 Continued

Indigenous category: Personal qualities	
Band/step	**Descriptors**
Band 2: Development needed	• The speaker's talk and behaviour point to questionable moral conduct as mentioned in the Band 1 descriptors but there is still room for the benefit of the doubt (MCR). • The speaker appears at times insincere, unhelpful, disinterested, and unapologetic even when they are in the wrong. Their goodwill is clearly lacking (DGW). • There is a noticeable absence of the spirit of collaboration and cooperation. The speaker can appear rather egocentric and dismissive of the needs of the group (CSQ). • Speaker's language appears rather flippant and dismissive, making others feel they are not respected (SRT).
Band 1: Intervention needed	• Showcases questionable moral values such as vileness, spitefulness, treacherousness, dishonesty, deceitfulness, connivance, untrustworthiness, disingenuousness or divisiveness (MCR). • The interaction clearly suggests that the speaker is an insincere, unhelpful, passive, indifferent or negative person (DGW). • The speaker appears to be very calculative, uncooperative, self-centred or self-serving with little to no regard for others. It is unrealistic to expect the speaker to sacrifice their own interest for the greater good (CSQ). • The speaker is extremely disrespectful to others. Language and behaviour are very flippant and dismissive (SRT).

5.2.3.7 Indigenous rating category: Reasoning skills

The third thematic analysis within the rating category *reasoning skills* identified three sub-categories: *providing solutions (PSS)*, *providing explanations and reasons (PER)* and lastly, *the structure of the talk (SOT)*, which used to be a separate linguistically-oriented category and later subsumed under reasoning skills (see Section 5.2.3.1). Each sub-category has five steps to differentiate different levels of performance. The *reasoning skills* category and its three sub-categories are presented in Table 33.

Table 33 Indigenous category: Reasoning skills

Indigenous category: Reasoning skills	
Band/step	**Descriptors**
Band 5: Exemplary	• When in face-threatening situations, adopts a solution-focused approach and supplies solutions that are effective, to-the-point and actionable. Demonstrates the ability to address the issue from different angles when necessary (PSS). • Showcases excellent reasoning skills and provides reasons or explanations that are sensible, defensible, logical, multi-angled, based on evidence and supported by examples (PER). • Solutions, reasons, and explanations provided are contextually and culturally appropriate (PER). • Talk is structured in a logical, sequential manner with clear signposting. Necessary sequential stages such as beginning, prefacing, development, main action, post action and closure are strung together in a cohesive manner (SOT).
Band 4: Good	• Can provide appropriate and effective solutions but solutions need more elaboration and detail to improve actionability and effectiveness (PSS). • Shows strong reasoning skills and supplies believable reasons, sometimes drawing on different perspectives. The examples and arguments used to support the reasons might need more substantiation (PER). • Talk is delivered in a logical manner with a basic structure guiding sequential stages such as beginning, development and closure (SOT).
Band 3: Average	• Shows the ability to provide solutions but solutions lack depth or might not be helpful to others in the long run (PSS). • Shows the ability to explain one's actions with reasons but the reasons employed are not the most readily acceptable ones. More elaboration on the reasons might be helpful (PER). • There is structure to talk with basic sequential stages in place, though the delivery of some stages has room for improvement (e.g., more details are needed in prefacing or closure is offered quite abruptly) (SOT).

Table 33 Continued

Indigenous category: Reasoning skills	
Band 2: Development needed	• Solutions are provided but are overly simplistic and ineffective. The solutions might even cause trouble to the speaker or others in the future (PSS). • Noticeable flaws in the reasons supplied: reasons might not be logical or believable enough; they can be one-angled and excessive, missing out on other aspects of the matter; they might also be superficial, poorly substantiated, or overly negative (PER). • There is reasoning but no solutions or vice versa (PER). • The structure of the talk is problematic with obvious flaws in some of the stages (e.g., prefacing is overly winding, or closure is missing) (SOT).
Band 1: Intervention needed	• Either does not provide any solution or proffers solutions that are problematic, poorly substantiated, inappropriate, ineffective, unhelpful, or impractical (PSS). • The reasons are either missing or unbelievable, unconvincing, insufficient, lacking in crucial details or appearing dismissive. The speaker shows poor reasoning skills (PER). • The reasons, solutions or explanations offered are contextually unacceptable or culturally inappropriate (PER). • The structure of talk is weak, resulting in ineffective delivery of information and unclear conveyance of intention. Basic sequential components are either missing or poorly designed (SOT).

5.2.3.8 Indigenous rating category: Social relations

The last indigenous rating category, *social relations*, has four sub-categories based on the third thematic analysis, the four sub-categories being *managing the speaker's roles (MSR)*, *attending to the interlocutor's roles (AIR)*, *attending to broader social relations (BSR)*, and *mediating multiple social roles (MMR)*. Table 34 explicates how the four sub-categories are defined and described in each of the five steps with sub-category abbreviations provided.

Table 34 Indigenous category: Social relations

Indigenous category: Social relations	
Band/step	**Descriptors**
Band 5: Exemplary	• Successfully builds social roles that are highly congruent with the roles designed for the speaker. Demonstrates strong ability to match one's behaviours to the expectations from one's roles and one's positioning in society (MSR). • Successfully attends to the interlocutor's social roles, the positioning of the interlocutor's roles and the interlocutor's role-congruent behaviours. Adopts behaviours and chooses language in a manner that is appropriate to one's interlocutor (AIR). • Demonstrates strong knowledge of the broader social relations that the speaker and interlocutor belong to. Showcases excellent ability to utilize such relations to achieve interactional goals (BSR). • Skilfully mediates the multiple, sometimes competing social roles that the speaker and the interlocutor possess and prioritizes primary roles when necessary (MMR)
Band 4: Good	• Can build relevant social roles in a believable manner though the details of the roles may be argued to be insufficient. The matching of one's roles with one's actions and one's positioning in society are overall appropriate, but inconsistencies can exist (MSR). • Demonstrates adequate awareness of the interlocutor's roles and interacts in a manner that is in general congruent with the interlocutor's roles, social positioning and the interlocutor's rights and entitlements (AIR). • Evinces sufficient awareness of and some application of broader social relations that are contextually appropriate (BSR). • Can balance the multiple roles that the speaker and the interlocutor have, though a more nuanced approach towards resolving role competition and prioritization would be desirable (MMR).
Band 3: Average	• The expected roles are overall demonstrated but not to an adequate level. The speaker can sometimes overdo or underdo the behaviours that are suitable for their roles. The speaker can in general position themselves correctly in society (MSR). • Displays some attention to the roles the interlocutor has, though there is inconsistency in adjusting to the interlocutor's roles, the interlocutor's expectations, or the interlocutor's social positioning (AIR). • Shows some understanding of broader membership relations. The utilization of them is passable but can be inadequate when more utilization could be desirable (BSR). • Can orient to the primary role of the speaker and the one of the interlocutor but other competing roles are largely unaddressed (MMR).

Table 34 Continued

Indigenous category: Social relations	
Band 2: Development needed	• The desired social roles are largely unbuilt, and the expected role-congruent behaviours are in general unfulfilled. The speaker's conduct mostly defies conventional expectations of their roles and their relative social positioning (MSR). • The speaker's choice of language and action overall lacks orientation to the interlocutor's roles and can violate the interlocutor's rights (AIR). • There is a noticeable miscalculation and misapplication of broader social relations. The speaker lacks knowledge of what appropriate broader social relations to draw on (BSR). • The primary roles of the speaker and the interlocutor and role competition are not sufficiently addressed (MMR).
Band 1: Intervention needed	• Failure at building the social roles as expected or building roles that are contextually inappropriate. The proposed behaviours are mismatched with the expected roles and social positioning of their roles (MSR). • Behaves in a manner that indicates a grave misunderstanding of or disregard for the interlocutor's roles, expectations, rights or social positionings. Conduct is unacceptable considering the interlocutor's social roles (AIR). • Serious neglect or mismanagement of social relations or the application of social relations that are unacceptable in the particular context. Relationships are damaged due to misunderstanding of broader social contexts (BSR). • Misidentification of primary roles and unacceptable conduct for either the speaker's or the interlocutor's primary roles. The multiplicity and competition of roles are unaddressed (MMR).

5.2.4 CA and MCA validation and the generation of exemplars

5.2.4.1 The rationale behind the CA and MCA validation of the scale

Section 5.2.3 offers a detailed account of how the indigenous IC rating scale was developed after three rounds of thematic analysis of DEs' comments. The resultant indigenous scale presents a sound representation of test-taker performance and is scaled based on DEs' perception. However, the scale itself is not sufficient to guarantee reliable ratings. To achieve an accurate rating, raters need to know how to connect the descriptive language in the descriptors in the scale to actual test-taker performance. In other words, effective rater training requires test developers to show raters how descriptors such as 'use supportive language to show one's care for other people' (Band 4 descriptor in the sub-category *being caring and supportive* BCS in the rating category *solidarity promotion* in Section 5.2.3.5) is actually achieved in interaction.

Another example is the descriptor 'successfully builds social roles(s) that are highly congruent with the role(s) designed for the speaker' in Band 5 in the sub-category *managing the speaker's social roles* (MSR) in the *social relations* rating category in Section 5.2.3.8. As discussed in Section 2.4.6 in the literature review, doing being specific social roles is a crucial component of IC. However, little previous research exists that looks specifically into how test-takers are doing being specific roles in a micro-analytic manner, drawing on their performance data. How such social roles are enacted in a moment-by-moment fashion is a process that requires explication for raters so that they are clear about what to focus on while assessing test-takers' performances. Such explication also has relevance for educators and language learners as they need to understand how 'doing being a doctor' or 'doing being a classmate' is achieved so as to be able to develop the competence to interact in a role-congruent manner.

In view of this, Section 5.2.4 reports on the qualitative analysis I conducted to connect DEs' comments and test-taker performance via Sequential-Categorial Analysis (CA and MCA) to illustrate how the rating categories are realized in empirical data. This process also generated empirical backing for the validation of the five rating categories as it served to verify if the features described in the rating scale and noted by DEs were actually traceable in test-taker discourse.

Another purpose of conducting Sequential-Categorial Analysis of test-taker discourse is to connect DEs' indigenous criteria with theories in CA and MCA. The indigenous IC scale developed from DEs' criteria in Section 5.2.3 is admittedly comprehensive and detailed in its coverage of the IC features considered crucial by DEs. However, due to the non-linguistic backgrounds of the DEs, their comments and the resultant indigenous rating scale lack theoretical insight and support. This is not an issue for the rating of the particular IC test in this book but could become problematic if the indigenous scale in Section 5.2.3 were to be used for other IC tests or other assessment contexts. Without a sound theoretical framework behind the rating scale, test developers are unable to assess if the abilities one scale assesses for a specific context can be extrapolated to other assessment tasks and contexts from a theoretical perspective. Drawing on my knowledge and training in CA and MCA, I could intuitively tell that the rating categories that DEs emphasized were connected with concepts in CA and MCA. Table 35 shows how DEs' atheoretical descriptors in the indigenous IC rating scale presented from Table 30 to Table 34 can find theoretical equivalents in CA and MCA concepts. Many of the CA and MCA concepts are discussed in Section 2.4.4 and 2.4.6 in the literature review in Chapter 2. To facilitate cross-referencing, footnotes of relevant literature on the CA and MCA concepts are provided in Table 35 for readers who would like to explore these concepts further.

The process of theorizing DEs' indigenous criteria at the stage of Table 35 is only tentative as the connection between DEs' descriptions and their CA and MCA counterparts is solely based on my training and perspective. In order to fully legitimize the theorization process, I need to inspect test-taker performance, in conjunction with DEs' comments and concepts in CA and MCA. This process can ascertain if CA and MCA theories can indeed capture DEs' descriptions. If the answer is affirmative, it shows that DEs' atheoretical comments can be theorized to provide theoretical structure to an otherwise atheoretical indigenous rating scale.

Table 35 Theorizing atheoretical descriptors in the indigenous rating scale

Atheoretical categories	Theorized categories	Atheoretical descriptors in the indigenous IC scale	Equivalent theoretical concepts
Conflict management	Disaffiliation control[4]	Conflict, face-threatening, interpersonal harmony, tact, blunt, aggressive	Disaffiliation, social solidarity, disaffiliative actions, dispreferred turn design
Solidarity promotion	Affiliation promotion[5]	Warm, supportive, face-saving, face-enhancing, empathy, putting oneself in others' shoes, positive, engaged	Pro-social, affiliative actions, preferred turn design, sharing of frames, empathetic unions, intersubjectivity
Reasoning skills	Reasoning[6]	Solution-focused, explanations, reasons, logical, different stages in talk	Remedies to interactional troubles, accounts, overall structural organization
Personal qualities	Morality[7]	Moral characters, virtues, goodwill, sincere, respectful, collectivistic, humble, team-spirit,	Protomorality (moral order 1, morality of interaction), cultural-specific morality (moral order 2, morality in interaction)
Social relations	Social role management[8]	Roles, appropriate behaviours to the roles, positions in society, social relations	Categories, activities/ predicates, relative hierarchical positioning, standardized relational pairs, membership categorization devices, duplicative organizations

4 Couper-Kuhlen (2012), Lindström & Sorjonen (2013) and Steensig & Drew (2008)
5 Goffman (1974), Haugh & Obana (2015), Hayashi (2013), Kim (2012)
6 Heritage (1988), Robinson (2013), Schegloff (2007b)
7 Bergmann (1998), Turnbull & Carpendale (2001), Turowetz & Maynard (2010)
8 Jayyusi (1984), Payne (1976), Sacks (1974), Stokoe (2012), Watson (1978)

5.2.4.2 The sample test task and the pilot test test-takers selected

To investigate if the rating categories are reflected in test-taker discourse, I transcribed pilot test test-takers' performance in CA fashion. Here performance excerpts on item 7 were selected. The performance excerpts also have the potential of serving as test-taker response exemplars for rater training, as discussed in Section 2.6.3. The task prompt and images from the prompt video are reproduced here for ease of reference.

Task prompt translated into English

'You recently went to a different city for work for six months and sub-let the apartment you rented to Wang Hao's son Wang Bin. Wang Hao is your best friend and has helped you a lot over the years. You also met Wang Bin when you visited Wang Hao in the past and had a good impression of him. Today the building manager of your apartment called you, telling you that recently there had been a lot of noise and loud music in your apartment late at night. Sometimes the neighbours also saw drunken youths coming in and out of your apartment. You want to discuss this with Wang Bin but since you are still away, you decide to talk to him via video chat.'

Figure 16 The test-taker sub-let their apartment

Figure 17 The test-taker receives a complaint call

Figure 18 The test-taker video-chats with the interlocutor

As the prompt makes clear, Wang Bin (the interlocutor) is making too much noise late at night and causing disturbance to the test-taker's neighbours. The prompt does not prescribe what social actions the test-taker needed to implement but most pilot test-takers oriented toward criticizing the interlocutor.

The performance exemplars selected mainly came from four pilot test-takers: L1-Chinese test-taker Xiaoxin and L2-Chinese test-takers Eric, Brian, and Hans. All four test-takers were from the 22 pilot test-takers (see Section 5.1.1.1). Pseudonyms were used. In terms of test-takers' backgrounds, Xiaoxin was an L1-Chinese speaker who grew up in China. Eric was an intermediate-proficiency L2-Chinese speaker who, at the time of undertaking the pilot IC

test, had never lived in China or any other Chinese-speaking regions. Similarly, Hans was a beginner-proficiency L2 speaker who had never resided in China. Brian, on the other hand, had lived in China for one year and had intermediate to upper-intermediate proficiency. Though it was safe to presuppose Xiaoxin's much stronger LC due to his L1-speaker status and much longer time of residence in the Chinese language community, Xiaoxin's performance on this particular item was perceived less favourably by the DEs than the ones from the three L2 speakers. It is also telling that although most of the L2 speakers selected here had never lived in China or only for a very limited time, their strong IC ensured their successful navigation through this task. The following is a presentation of the exemplars that unpack DEs' indigenous criteria and can serve as CA-informed rater training materials.

In total nine exemplars are presented here. Exemplar 1 to Exemplar 3 represent all the five rating categories in test-taker discoursal data and connect DEs' etic comments with my CA inspection of the data. Exemplar 4 to Exemplar 9 focus more specifically on the last rating category, *social relations* (see Section 5.2.3.8), due to the limited attention on the categorial aspect of interaction received in IC research so far. DEs' etic comments for relevant exemplars are provided for Exemplar 1 to Exemplar 3, but not Exemplar 4 to Exemplar 9 due to space restrictions. The aim of the analyses of exemplars is to demonstrate that DEs' rating categories can be located and explained in test-takers' discourse. Concurrently, I strive to show the value-add of Sequential-Categorial Analysis through its power in unpacking DEs' oft-vague comments, defining the test construct with more precision, and generating exemplar-based rater training materials to standardize raters' understanding of the test construct. The CA transcription followed the Jefferson style (Jefferson, 2004) and the glossing followed the convention in analysing Chinese CA data (Wu, 2003).

5.2.4.3 Theorizing conflict management and social relations

Below is a presentation of Exemplar 1 which showcases how the indigenous rating categories *conflict management* (Section 5.2.3.4) and *social relations* (Section 5.2.3.8) can be theoretically expanded through Sequential-Categorial Analysis. Exemplar 1 is based on the performance of test-taker Xiaoxin, an L1-Chinese speaker, with a trained L1-Chinese interlocutor role-playing Wang Bin, who was supposed to be the son of Xiaoxin's friend Wang Hao.

Exemplar 1 Xiaoxin: 'you're bringing me contempt'

1 I E:::H .h jiushi (.) shushu jue >hui bu hui< juede E:::R
PRT just uncle feel can N can feel PRT
'Ehh uncle, do you thi(nk), think if it's possible Errrr'

2 I >shi bu shi< linju tai: <mi:nggan> le a
be N be neighbour too sensitive ASP PRT
'if it is the case the neighbours are being too sensitive here?'

3 (1.4)

4 X E:::N ↑H::N JIUshi .hh buguan linju duo minggan (0.1)
PRT PRT just no matter neighbour much sensitive
'Enn Hmm well, no matter how sensitive the neighbours are'
5 X zhe ge shi WO jia= wo huiqu deshihou haishi yao zaodao=
this C be my home I return when still will encounter
'this is my home. When I come back I will still be met with'

6 X =linju de baiyan de= jiushi (0.2)
neighbour ASSC contempt NOM just
'contempt from the neighbours, that's why'

7 X ni yao wei wo zhaoxiang yi dian-
you need for I think one C
'you should think more from my perspective'

8 I (0.4) .tch (0.5) .hh E::::N (0.5)
PRT
'Enn'

```
9   I   <na:   wo hui>  shaowei:    zhuyi           yixia        ba:
         then  I  can    a little    pay attention   a bit        PRT
        'in that case, I will be a bit more careful then'

10  I   (0.1)   ((ci- abandoned))    (0.3) E:::M (1.6)
                                           PRT
        'Emmm'

11      (1.6)

12  I   [ xi-]  <xing=  >buhao-yisi   a-   gei shushu tian  mafan   le-<
                 OK      embarrassed  PRT  to  uncle   add  trouble PRT
        'All, all right then, I'm sorry for causing trouble for you'

13  X   [ o  ]
         PRT
        'oh'
```

Note. pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and X for Xiaoxin

Amongst the four DE raters, DE25, DE26, DE27 and DE28, who listened to the 22 test-taker performances on this item (see Table 25 for the matching between DEs and the test items), DE25 and DE28 explicitly commented on Xiaoxin's talk from lines 4–7 as being problematic. Both DE25 and DE28 loosely orthographically transcribed lines 4–7 when listening to the recording and noted in their written comments respectively 'as a senior member (Xiaoxin), inappropriate' and 'this line is inappropriate because it makes the other person not sure what to say in response'. During the focus group interviews, I utilized stimulated recalls to prompt DE25 and DE28 to further elaborate their thoughts. Below is a translated summary of the interview transcripts of DE25's and DE28's comments in relation to Exemplar 1. I used the line numbers in Exemplar 1 to refer to the timestamps in DEs' comments for ease of reference. I annotated DEs' comments in brackets when the referents were unclear.

> DE25: It (lines 5–7) is not exactly a threat, but it's rather harsh. The expression *wei wo zhaoxiang* 'think more from my perspective' makes me feel rather uncomfortable. Another consideration is Wang Bin (interlocutor) is his (Xiaoxin) friend's son, there is this relationship here… it (the way the test-taker talks) just doesn't feel right.

DE28: Lines 5–7 makes me feel quite uncomfortable as well, for a start. He (Xiaoxin) makes Wang Bin (interlocutor) unable to respond. You can see that from the other person's (interlocutor) response. He doesn't know what to say. The underlying message (line 5) is 'this is not your place; you can't do whatever you want to do'. The thing is he (Xiaoxin) is a *zhangbei* (senior member of society), and he is good friends with the other person (interlocutor)'s father. He (Xiaoxin) has two roles here. The way he (Xiaoxin) talks sounds like he is getting into a fight, not befitting his role. You (Xiaoxin) can offer criticism or suggestion, but you can't make the other person (the interlocutor) lose face.

To sum up, DE25 and DE28 raised three issues with Xiaoxin's conduct in Exemplar 1: lines 5–7 were overly critical and unsettling to the raters, Xiaoxin did not talk in a manner that was congruent with his social role in the task, and he disrupted the flow of interaction by rendering the interlocutor unable to respond. Though DEs' written comments and interview discussions successfully identified the snippets that were problematic and elaborated their rationale for problematizing them, their rationale, as reproduced above, was still somewhat vague and impressionistic in its description, due to its a-theoretical nature. Now I will employ Sequential-Categorial Analysis to analyse Exemplar 1 in order to unpack DEs' sense-making of their 'obdurate world of everyday life' (Turowetz & Maynard, 2010, p. 505), connecting test-taker discourse with the two indigenous rating categories of relevance here: *conflict management* and *social relations*.

Lines 1–2 from the interlocutor are in response to Xiaoxin's news-telling of the complaints from the neighbours, which was a veiled criticism from Xiaoxin of the interlocutor for causing annoyance to Xiaoxin's neighbours. The interlocutor was trained to push back the test-takers' criticism so as to elicit more display of IC from the test-takers. Here the interlocutor disagrees with Xiaoxin by offering a different assessment of the situation. The word *minggan* 'sensitive' in line 2 deflects neighbours' grievances to themselves for being too sensitive. The two evidential markers *huibuhui* 'if it is possible' and *shibushi* 'if it is the case' epistemically downgrade the interlocutor's claim, softening the strain on solidarity engendered by the dispreferred nature of disagreement.

Contrary to the interlocutor's careful management of disaffiliation, Xiaoxin, after a 1.4-second gap in line 3, delivers a bald on-record criticizing action, which is not only disaffiliative but also borderline threatening as DE25 noted. In line 4, Xiaoxin first uses an extreme case formulation (Pomerantz, 1986) *buguan* 'no matter', discrediting the legitimacy of the interlocutor's candidate explanation and forestalling any further rationalization from the interlocutor.

Xiaoxin's on-record criticism centres on the word *baiyan* 'contempt' in line 6, a highly affectively disaffiliative and morally implicative expression that upgrades a neighbourly tiff to the harrowing prospect of Xiaoxin being treated as a social pariah by his neighbours upon return, due to the interlocutor's misdemeanour. In terms of Xiaoxin's overall delivery, lines 4 to 7 are extremely smoothly delivered with no gaps between sentences. At a transition relevant place after *de* in line 6, Xiaoxin's use of *jiushi* 'just' successfully holds the turn, and after a 0.2-second gap, Xiaoxin reinforces his criticism by spelling out to the interlocutor that he should have been more considerate. If rated against traditional Leveltian speaking/LC rubrics, Xiaoxin's performance would score high in terms of lexical range, syntactic complexity and fluency. However, from the perspective of talking/IC and social solidarity, Xiaoxin's delivery of criticism lacks all the mitigating devices commonly seen in dispreference structures. It is unsurprising that DEs flagged Xiaoxin's lines 4–7 against the indigenous rating category *conflict management* (Section 5.2.3.4).

The second concern from DEs related to the incongruency between Xiaoxin's talk and his social role, which is pertinent to the *social relations* indigenous rating category (Section 5.2.3.8). If uninformed of the context in the role-play prompt, one can read Xiaoxin's lines 4–7 as an utterance recipient-designed to a 'tenant', to an 'acquaintance' but not to your 'best friend's son'. The emphasized *wo jia* 'my home' in line 5 is an explicit statement of Xiaoxin's ownership of the apartment, establishing his category of the 'houseowner'. Operating within this categorization, Xiaoxin evokes the rights and entitlements bound to the 'houseowner' category such as 'asserting the owner's overriding right to the house', 'requesting the tenants to obey the houseowner's rules' and 'expecting the tenants to show consideration to the houseowner's needs for neighbourly amicability', all of which are used as grounds by Xiaoxin to reprimand the interlocutor in lines 4–7. However, as both DE25 and DE28 commented, Xiaoxin fails to orient to other crucial social roles made relevant by the task prompt. Instead of licensing a hierarchical relational pair of 'houseowner' and 'tenant', Xiaoxin should have attended more to the interlocutor's category of being 'his best friend's son' and his category of being a 'more senior member in society', both of which were implicated in the task prompt. The lack of orientation to such categories made the DEs consider Xiaoxin's talk misaligned with the social roles of his interlocutor and incongruent with the social roles Xiaoxin was supposed to assume, translating to unimpressive feedback in relation to the rating category *social relations*. More Sequential-Categorial Analysis of the categorial dimension of this task can be found in Exemplar 4 to Exemplar 9.

The last issue DEs took with Xiaoxin's conduct is the interactional disruption his lines engendered for the interlocutor, which is painfully palpable when the speech is analysed using Sequential-Categorial Analysis. The repeated pauses in line 8, the prolonged hesitation markers *EN* and *EM* in lines 8 and 10, the started but abandoned turn in line 10, the distressing 1.6-second-gap in line 11 and the rush-through in line 12 are all demonstrative of the interlocutor's struggle in formulating a response to Xiaoxin's sharp criticism. Looking at the exemplar as a whole, the work of maintaining social solidarity has been on the interlocutor's shoulders alone, evidenced by the interlocutor's work at mitigating disaffiliation in lines 1–2, the use of mitigators *shaowei* 'a little' and *yixia* 'a bit' in line 9, the deployment of the epistemic downgrader *ba* in line 9 (Kendrick, 2018), and the proffering of an apology in line 12. Xiaoxin's lines, on the contrary, have predominantly served as a disruptive force to interaction. It is therefore unsurprising that, despite Xiaoxin's L1-speaker advantage and strong LC, his performance on this item was viewed negatively by DEs on their indigenous criteria of interactional success.

In summary, Exemplar 1 shows that the indigenous rating categories *conflict management* and *social relations* could be theorized using CA and MCA concepts and such theorization was validated in test-takers' discourse. The two indigenous categories were then renamed *disaffiliation control* and *social role management* to better capture the theoretical import from CA and MCA. The next exemplar shows how the *solidarity promotion* and *reasoning skills* indigenous categories can be similarly theorized.

5.2.4.4 Theorizing solidarity promotion and reasoning skills

Exemplar 2 below presents an interaction between the trained interlocutor and an L2-Chinese pilot test-taker, Eric (see Section 5.2.4.2 for Eric's profile). The stage of the interaction in Exemplar 2 was similar to the one in Exemplar 1 but we can see Eric handled the interaction in a different manner compared to Xiaoxin. Exemplar 2 also showcases how the indigenous criteria *solidarity promotion* and *reasoning skills* can be theorized via Sequential-Categorial Analysis.

Exemplar 2 Eric: 'can't choose thy neighbour'

1 I hui bu: hui: shi linju tamen
can N can be neighbour they
'is it possible that the neighbours themselves'

2 I tai <minggan> le a: ∘ni juede∘
too sensitive ASP PRT you think
'I mean, do you think they are being too sensitive here'

3 E ↓TAI mi:ng↓gan ↑E::M yexu shi zhe yang e-
Too sensitive PRT maybe be this manner PRT
'too sensitive, Emmm, maybe it is like that'

4 E wo=wo >youdian< tongyi= danshi:: A:: (0.9)
I I a bit agree but PRT
'I, I agree with you somewhat, but, ahhh'

5 E women bu: ne::ng xuanze women de:: linju
We N can choose our ASSC neighbours
'the thing is we cannot choose our neighbours'

6 E (0.3) .hh ye:: ye you: ↑gui↓lv ba hm=
also also have order PRT PRT
'this is just how things are, hmm'

7 I =∘em::∘ hao: na wo yihou hui zhuyi de=
PRT good then I after ASP pay attention NOM
'Emm, ok, in that case I will be more careful in the future'

8 I =buhao-yisi e [mafan shushu] le
Embarrassed PRT trouble uncle ASP
'I'm sorry for trouble you'

9 E [en= en]
PRT PRT
'hm,hm'

```
10  I  nin   hai    zai  [chuchai
       You   still  at    business trip
       'especially when you are still on a business trip'

11  E                    [AH:: mei wenti
                          PRT   N   problem
       'Ahh, that's no problem at all'
```

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and E for Eric

Mirroring Exemplar 1, Exemplar 2 starts with the interlocutor shifting the blame to the neighbours for being excessively sensitive. Different from the scathing comments on Xiaoxin's performance, DE raters viewed Eric's turns very positively. Both DE26 and DE27 transcribed Eric's lines 3–4 in their written comments and wrote 'being very measured' and 'accepting, understanding, listening'. They further expanded their observations in the focus group interviews.

> DE26: When the kid (interlocutor) asks if the neighbours are being too sensitive, I think his (Eric's) response deserves extra points. What he says (lines 3–4) shows he is not negating the kid's idea; he is maintain an accepting attitude. No matter how much he (Eric) actually agrees with him (interlocutor) in line 4, at least he conveys an affirmative attitude, he (Eric) doesn't contradict him, but addresses the issue from a different angle in line 5, which is also a really good reason.

> DE27: What he (Eric) says in line 4 maybe is a lie, maybe he doesn't really agree, but he considers the kid's feelings, that's why he says that. He also says we can't choose our neighbours (line 5). Whether the neighbours are truly being oversensitive or not, we still can't pick and choose them. I think here he also deserves extra points. What he says is empathetic and full of reason at the same time.

The above interview summaries show that DE26 and DE27 thought highly of how Eric strengthened solidarity with the interlocutor in lines 3–4 and how he proffered legitimate reasons to substantiate his claim in line 5, corresponding to the *solidarity promotion* and *reasoning skills* categories in the indigenous rating scale. Now let us turn to the CA transcript to unpack DEs' comments and inspect emically the interactional methods Eric mobilized to make his interaction so recognizably different from Xiaoxin's.

Similar to lines 1–2 in Exemplar 1, lines 1–2 in Exemplar 2 are semi-prescribed for the interlocutor to prompt the test-taker to produce more language. We can see the same evidential marker *hui bu hui* 'is it possible' and the same stress on the word *minggan* 'sensitive'. The epistemic downgrader *ni juede* 'do you think' is produced *sotto voce*, hedging the claim the interlocutor makes regarding Eric's knowledge (Lim, 2011) and inviting Eric to assess the interlocutor's candidate explanation.

Faced with a similar first-pair-part from the interlocutor, Eric's second-pair-part disagreement, spanning from lines 3 to 6, differs greatly from Xiaoxin's. Although it is disaffiliating in nature just like Xiaoxin's, Eric prefaces his disagreement in lines 5–6 with an agreement from lines 3–4, which is a well-documented format that mitigates disaffiliation (Pomerantz, 1984). Eric's agreement consists of three components, each of which implements a different affiliative move. In line 3 he first produces a partial repeat of *tai minggan* 'too sensitive', as a receipt and confirmation of the interlocutor's claim, demonstrating attentive listening and evidencing intersubjectivity, which was noted by the DEs. The ensuing *yexu shi zheyang* 'maybe it is like that' affiliates with the interlocutor epistemically. Given that the interlocutor is only living in Eric's apartment temporarily, Eric is putatively the more knowledgeable of the two in terms of his neighbours' dispositions. Eric could have oriented to his epistemic advantage and refuted the interlocutor's claim, but instead, he treats the interlocutor as having equal access to this particular domain of knowledge, which promotes solidarity. In line 4 Eric's use of *tongyi* 'agree' further affiliates with the interlocutor, but this time it is affectively affiliating instead of epistemically affiliating. Though the mitigators *yexu* 'maybe' in line 3 and *youdian* 'somewhat' in line 4 are hearable as foreshadowing disagreement, the partial repeat, the epistemic and the affective matchings have worked to buffer the disaffiliative impact of disagreement. The rich affiliation promotion Eric has employed merits the positive 'good at listening', 'not negating other's opinions' and 'showing agreement to be empathetic' comments from the DEs.

Moving on to the disagreement component in Eric's utterance, this is where he demonstrates his strong reasoning skills and, as DE26 put it, 'addresses the issue from a different angle'. Eric's disagreement starts after *danshi* 'but' in line 4 and is followed by a prolonged pre-disagreement hesitation marker *A::* before the disagreement proper is launched in lines 5 and 6. The temporal positioning of the disagreement component showcases Eric's strong reasoning competence as he places it after the three affiliative moves as discussed above, which prevents his disagreeing account from being heard as overly critical (Heritage, 1988). The disagreement proper in line 5 is the highlight of Eric's reasoning as the disagreeing action is realized not through an on-record one such as Xiaoxin's

criticizing but through an off-record explanation based on commonsensical reasoning. The explanation that one cannot choose their neighbours in line 5 orients to members' tacit, ordinary understanding of social order, the normative nature of which renders its legitimacy irrefutable. The irrefutability of Eric's account can also be observed via his use of *women* 'we' in line 5, suggesting that what he is about to say is an intersubjectively shared fact. It is worth noting that Eric's account is a 'no-blame' account. Similar to inability and no-fault accounts (Drew, 1984; Heritage, 1984),here I define a no-blame account as appealing to reasons or explanations outside the control of the interactants, refraining from putting the blame on the interactants and hence preserving social solidarity. In line 6 Eric further underscores the legitimacy of his account with the word *guilv* 'order', which highlights the commonsensical nature of his account. This account is moderated by a sentence-final particle *ba* in line 6, which downgrades Eric's epistemic position and affiliates with the interlocutor epistemically (Kendrick, 2018). Though linguistically *guilv* is an infelicitous word choice and indexes Eric's restricted lexical repertoire, line 6 progresses the interaction and did not raise alarms for the DEs. This points to the separation between traditional measures of speaking/LC and measures of talking/IC.

Moving beyond the lines commented by DEs and looking at the interaction as a whole after Sequential-Categorial Analysis, the praiseworthy work by Eric on affiliation promotion and reasoning has paved the way to smooth interaction between him and the interlocutor in the ensuing talk from lines 7–12. After a subdued *em*, the interlocutor promptly supplies an agreement token *hao* 'ok', promises he will be more careful in the future in line 7 and apologizes for causing trouble for Eric in line 8. Line 10 demonstrates the interlocutor's attempt at preserving solidarity by thinking from Eric's perspective. The interlocutor's pro-social moves are generously reciprocated by Erics' aligning token *en en* 'hm hm' in line 10 and affiliative *mei wenti* 'no problem' in line 11, which overlaps with the interlocutor's solidarity bid in line 10 and reinforces affiliation by dispelling any concern the interlocutor might have had at that point.

In summary, Exemplar 2 demonstrates a very different type of interaction between test-taker and interlocutor. Instead of relying on the interlocutor solely for preserving solidarity as observed in Exemplar 1, here both Eric and the interlocutor adopt various pro-social moves to collaboratively maintain solidarity. Eric's demonstration of affiliation promotion and reasoning skills are particularly commendable, making this snippet a useful exemplar to illustrate desirable performances in rater training. This exemplar also shows that the indigenous rating categories *solidarity promotion* and *reasoning skills* can be formally theorized as *affiliation promotion* and *reasoning* in CA parlance. The

next exemplar illustrates how the indigenous rating category *personal qualities* can be theorized.

5.2.4.5 Theorizing personal qualities

Exemplar 3 is also from Eric but centres on his performance for the *personal qualities* rating category in the indigenous rating scale.

Exemplar 3 Eric: 'we need to respect them'

```
1  E  hao    NA:: (0.2)  jiushi  ni   yinggai  zhidao  yinwei:::=
      Good   that        just    you  should   know    because
      'Ok, in that case, it's just that you should know, because'

2  E  =a:::  women  de   linju:: (0.5) A:::  you xie:: (0.1)
       PRT   our    ASSC neighbour     PRT   be  PL
      'ahhh, among our neighbours, ahhhh, there are some'

3  E  A:: (0.5) nianji daren     en::   bijiao      da   de   ren
      PRT       age    grown-up  PRT    relatively  big  NOM  person
      'ahhh, older grown-ups, hmmm, relatively older people'

4  E  e:   ye    you::  e::  (0.4) jiating  gen   haizi:   men::
      PRT  also  have   PRT        family   and   kid      PL
      'Errr, they also, errr, have families and children'

5  E  zhe   yang   de    ne:=
      this  manner  NOM   PRT
      'things like that'

6  E  =women  YINGGAI::  a:      zunzhong      tamen  (0.2)
       we     should     PRT     respect       them
      'we should respect them'

7  I  ∘em:∘=
       PRT
      'em:'
```

8 E =ah: bu:↑yao:: (1.8) ah zai ye↓li zuo::
PRT N PRT at night do
'ahh, when ahhh, it is night time, we shouldn't do'

9 E tai DA she:n zhe yang de
too big sound this manner NOM
'anything that is too loud'

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and E for Eric

In the interview, DE25 and DE26 commented on Eric's character as a member of society, which implicitly oriented to Eric's management of the moral order of everyday interaction.

> R25: I think his (Eric's) description in line 6 is quite placid, moderate, and aligned with what people would normally think.

> R26: I agree with what R25 just said. Although he (Eric) is a foreigner, his talk is very to the point. For example, he talks about neighbours' families and kids at 1 minute and 15 seconds (line 4), I made a note there. He then says 'we should respect them' (line 6). What he is saying here is very spot-on, and he is not criticizing the other person (interlocutor). It is the ordinary, normal behaviour you would expect from anyone, such as not making too much noise at night.

Though R25 and R26 did not explicitly refer to the concept of *morality of interaction*, their comments on the 'what people would normally think' and 'ordinary, normal behaviour' attested to their orientation to everyday social order as moral order and what a moral being in society implicates (Bergmann, 1998). They were impressed by how Eric, despite his L2 speaker status, was able to appeal to the normative moral expectations of members in a shared society to substantiate his argument, without resorting to disaffiliative actions such as criticizing as Xiaoxin did in Exemplar 1. Now let us inspect the CA transcript to investigate how moral order is realized in this exemplar.

Eric's moral work starts with the modal auxiliary *yinggai* 'should' in line 1. *Yinggai,* coupled with *zhidao* 'know' in line 1, projects what Eric is about to say is presupposed as an intersubjective fact shared by the interlocutor. By aligning the interlocutor's epistemic status with his own and asserting the equalness of their knowledge domain, Eric foregrounds the knowingness on the interlocutor's

side and intimates the moral incontrovertibility of his upcoming claim. Following the morally projective line 1, Eric details the specifics of his neighbours' living situations such as their advanced age and their co-inhabitation with families and children from lines 2 and 5. This description does not exist in the task prompt and is constructed *in situ* by Eric. Such a creative characterization of neighbours serves as the bedrock of Eric's moral work. In line 6 Eric launches his moral claim on this matter. He first prosodically upgrades the same modal auxiliary *yinggai* 'should' as in line 1. Then follows the focal point of Eric's argument, *zhunzhong* 'respect', which moralizes this neighbourly incident and explicates what Eric considers to be morally imperative in this context. A mutually respectful relationship between neighbours is a normative, natural relationship to which moral members of society should and would naturally orient. This relationship can be strained by morally sanctionable conduct such as making too much noise at night, which is what the interlocutor is doing and is encapsulated by Eric in lines 8–9. By making explicit the moral order and moral implications of this incident, Eric holds himself to his expectation of the behaviours of a moral member. By emphasizing the intersubjective sharedness of this moral order, Eric subjects the interlocutor to the same moral standards and expectations, which as DE25 and DE26 commented, are 'normal' and 'ordinary'. Understandably there are similar linguistic restrictions in Eric's utterance such as the self-repair on the awkward construction *nianji daren* 'older grown-ups' in line 3. However, the moral work undertaken in this snippet corroborates Eric's strong IC in that he demonstrated the ability to draw on everyday morally implicative moments to strengthen his argument and attain interactional success. From a theoretical perspective, the comments DEs made on Eric's performance in relation to his personal qualities can be reinterpreted as Eric's attempts at preserving the moral order of interaction. The indigenous rating category, *personal qualities*, therefore, lends itself to be theorized as *morality* from the perspective of CA and MCA.

5.2.4.6 Address terms in social role management

The above three exemplars, Exemplar 1 to Exemplar 3, provide evidence on how the five indigenous categories, *conflict management*, *reasoning skills*, *solidarity promotion*, *personal qualities,* and *social relations*, can be theorized into categories informed by CA and MCA concepts, namely, *disaffiliation control*, *reasoning*, *affiliation promotion*, *morality* and *social role management*. The readers can refer to Table 35 for a detailed presentation of how DEs' atheoretical indigenous rating categories are connected with the CA- and MCA-informed theorized IC rating categories.

The three exemplars, Exemplar 1 to Exemplar 3, due to their brevity, only shed light on certain aspects of the five indigenous rating categories. Considering the relatively lesser focus on the categorial dimension (MCA) of IC compared to the sequential dimension (CA), here more exemplars from the same task are offered to further explicate the *social role management* rating category as the analysis of this category relies heavily on concepts in MCA.

Exemplar 4, featuring Brian, a different L2-Chinese test-taker, illustrates how social roles can be initially established by address terms.

Exemplar 4 Brian: 'call me shushu'

```
1   B      .tch (0.2) Wang Bin          [ni hāo:
                      (full name)       hello
           '.tch (0.2) Wang Bin         [hello'

2   I                                   [Ai
                                        PRT
                                        ['Hi'

3   I  →   (.) Ai         Brian             shushu  hao
               PRT     (first name)         uncle   good
           '(.) Hi Uncle Brian,hello'

4          (0.9)

5   B      .tch (0.2) .tch Zui↑jin   zenme   ↓YANG?
                        Lately    how     condition
           '.tch (0.2) .tch how have you ↓been? la↑tely'

6          (0.4)

7   I      AAA:    >∘zuijin    haihao∘<        jiu      xuexi  >∘shenme de∘<
           PRT       lately    pretty good     just     study    what   NOM
           'AHH: >∘Pretty good lately∘< just studying >∘and the sort∘<'
```

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and B for Brian

In line 1 Brian launches a greeting sequence, addressing the interlocutor (I in the transcript) by the full name of his role Wang Bin. The interlocutor overlaps with Brian's talk, produces an acknowledgement token in line 2 and returns

the greeting in line 3, suggesting that Brian's greeting is formatted in an easily recognizable manner. What is of analytical interest here is that the interlocutor addresses Brian using his English name, followed by a Chinese kinship term *shushu* 'uncle'. On the interlocutor's prompt it was not specified how the test-taker should be addressed but here the L1-Chinese interlocutor draws on his member's knowledge and decides to address Brian with a kinship term *in situ* and *in vivo*. *Shushu* in Chinese is not reserved only for blood relatives but is an inference-rich address term that carries specific duties, obligations, and expectations. A discussion on categories in MCA and their concurrent activities/predicates can be found in Section 2.4.6. There are ample cases in the pilot test-taker dataset where test-takers self-categorize as either *shushu* 'uncle' or *ayi* 'auntie'. Two additional examples are presented here in Exemplar 5 and Exemplar 6 for illustrative purposes.

Exemplar 5 Test-taker: 'I'm shushu'

1	T	Wo xiangxin (0.4) E: ni ye shi ↑ge:: (0.4)	
		I believe PRT you also be C	
		'I believe (0.4) Er: you are also a (0.4)'	
2		DONGSHI de xiao hai (0.5)	
		understanding NOM small kid	
		'kid that is UNDERSTANDING'	
4		E: wo xiangxin ni keyi zuo hao de	
		PRT I believe you can do good NOM	
		'Er: I believe you can behave well.'	
5	→	∘E∘ shushu wo ye (0.2) wo ∘ne∘, ye hen xinren ni (0.4)	
		PRT uncle I also I PRT also very trust you	
		'As your uncle I also (0.2), I also really trust you (0.4)'	
6		Hm [hao ba].	
		PRT good PRT	
		'Hm, [ok?]'	

```
7  I            [.tch    ]

          (0.4)

8  I     ↑Ou:          [shushu]
          PRT           uncle
         '↑Oh: but      [uncle  ]'

9  T                    [Hm     ]
                         PRT
                        [Hm      ]
```

Note. T stands for test-taker and I for interlocutor.

Exemplar 6 Test-taker: 'I'm ayi'

```
1   I    Oh::=
         [PRT]
         'Oh:::='

2   T →  =Ayi    juede  ni men xiao  nianqin: na (.) keyi   jiushi (0.3)
         =auntie think you PL  small young    PRT     can   just
         '=Auntie thinks young people: like you(.)you can just (0.3)'

3        jiushi OUER         yule  hai  yixia (0.1) dou meiyou wenti=
         just  occasionally  relax fun  once        all no     question
         'just OCCASIONALLY relax and chill a bit(0.1)that's all fine= '

4        =danshi ne:   ye     buyao jiushi
          but    PRT   also   don't  just
         '=but just don't'
```

Note. T stands for test-taker and I for interlocutor.

Going back to Brian's performances in Exemplar 4, instead of resisting the interlocutor's attempt at categorizing, Brian starts in line 5 with a question enquiring about the interlocutor's wellbeing. Brian's implicit sanctioning of the category *shushu* shows that the interlocutor's activity is permissible and pre-existent in both parties' shared category knowledge. By establishing intersubjectivity, it also licenses Brian to exploit the inferences bestowed by the *shushu* category.

5.2.4.7 Categories and predicates

Building on the analysis in Exemplar 4, Exemplar 7 shows how pilot test-taker Brian establishes his social role vis-à-vis the role of his interlocutor, drawing on the categories and category-bound activities inherent in the social roles in item 7.

Exemplar 7 Brian: 'I want to be nice to you'

```
1    B     >NI   ZHIDAO  WO  shi< (1.7) ni     BA,
           You  know    I   be         your   father
           '>YOU KNOW I am< (1.7) your DAD'

2    B     jiu (.)  >wo de     hao    pengyou<
           just      I  ASSOC  good   friend
           just (.) >is my good friend<'

3          (0.4)

4    I     Dui=dui=∘dui∘
           Yes  yes yes
           'Yes=yes=∘yes∘'

5          (0.7)
6    B →   Aaa:::   (0.3) Suoyi (.)    wo   xiāng  ↓DUI      ni     hao
           PRT            therefore    I    want   towards   you    good
           'Ahh:::   (0.3) Therefore (.) I want to be good ↓TOWARDS you'

7    B     (0.3)Wo ye     ↑XI↑WANG     ni: (0.1)   ni   ye:
                I  also   hope         you        you   also
           '(0.3) I also ↑HOPE  you: (0.1) you also'

8          (0.9)
```

```
9    B →  dui        WO:    hao.
          towards    I      good
          'are good towards ME'

10         (0.4)

11   I     Ye::     ↓en↑hen      (0.3) ye?
           PRT      PRT  PRT            PRT
           'Yes::    ↓um↑hum (0.3) Yeah?'
```

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and B for Brian

A few turns after Exemplar 4, Exemplar 7 takes place where Brian makes use of his *shushu* category when he criticizes the interlocutor. Brian first makes an explicit attempt at self-categorization, abandons it, topicalizes the test-taker's father, and establishes a standardized relational pair between test-taker's father and himself from lines 1–2. A standardized relational pair is a pair of categories that carry mutual obligations and responsibilities (Stokoe, 2012), such as patient-doctor, salesperson-customer and friend-friend. By categorizing the interlocutor's father as his *good friend*, Brian also self-categorizes as a *good friend* to the interlocutor's father, as friendship cannot be claimed one-sided. Brian's relational pair and self-categorization are quickly endorsed by the interlocutor with repeated acceptance tokens in line 4. When we combine the categorization work of 'shushu' in Exemplar 4 and the relational pair in Exemplar 7, we can see how the three categories, *father*, *son,* and *uncle* emerge as duplicatively organized (Francis & Hart, 1997). This makes explicit a commonsensical practice in Chinese society: if a male Chinese speaker is on close terms with another male Chinese speaker of similar age, the son of one of the speakers is bound to call the other speaker *shushu* 'uncle', despite there being no blood connection between the son and the other speaker. Such knowledge is neither explicated in nor mandated by the scenario script for test-takers or the prompt for L1-Chinese interlocutors. It is both parties' members' knowledge of their respective social roles that make them co-construct this duplicative organization.

Having examined the category-building work from both Brian and the interlocutor, now let us look at how Brian makes his laborious categorization work for him. Categories are not void concepts. Sacks first noted that categories have activities bound to them, such as in his famous example the baby (category) cried (activity) and the mommy (category) picked it up (activity) (Sacks, 1974). Subsequent work on MCA has further substantiated the activities (also called

predicates) tied to categories, such as entitlements, duties and knowledge (Jayyusi, 1984; Payne, 1976; Watson, 1978). In Exemplar 7, after sanctioning his category as *shushu* and the interlocutor's category as *close friend's son*, in line 6 Brian makes explicit the obligations he perceives to have been engendered by their relational categories: 'I/*shushu* wanted to be good towards you/*close friend's son*' and 'I/*shushu* hope you/*close friend's son* are also good towards me/*shushu*'. Note Brian's choice of verbs here. When he explicates his moral obligation as a *shushu*, he uses *xiang* 'want' in line 6, a word indexing stronger agency. When he describes the obligation of his friend's son, he chooses *xi wang* 'hope' in line 7, a word that hedges the proposition. Similarly, the word *ye* 'also' offers a softening effect as it emphasizes reciprocity.

In line 11, the interlocutor first issues an affiliative token *ye* 'yes' with a downward intonation, licensing Brian's category inferences and acknowledging the mutual obligations that Brian purposely establishes. The subsequent rising and falling intonation of the particles *en hen* 'um hum' is particularly intriguing here. *En hen* 'um hum' suggests that the interlocutor predicts that there is more to come from Brian since Brian's exposition of mutual obligations is knowledge that already exists in their shared category knowledge. The fact that Brian has painstakingly spelled out such details of their category predicates can only implicate one thing: Brian wants to make use of such predicates. Indeed, though not reproduced here, Brian's 'I wanted to be good towards you' later paves the way into his more attentive enquiries of what the interlocutor does at night, what kind of people he brings home and whether they have been drinking irresponsibly. Brian's 'I hope you will be good to me too' provides grounds for Brian's criticizing of the interlocutor's behaviour as what the interlocutor does obviously is not 'being good' to *shushu*. Everything takes on a different hue after the meticulous category groundwork Brian has laid for his criticizing action. Though what the interlocutor has done is irresponsible in itself as in disturbing the neighbours, Brian adds ammunition to his criticism by evoking the moral obligations rooted in their categories. Therefore, what the interlocutor has done has thus become not just irresponsible, but also immoral.

5.2.4.8 Beginner L2-speakers' category knowledge

The previous exemplars on Brian demonstrate that pilot test-takers did not solely focus on launching the intended social action implied in the task scenario. They contextualized actions in a manner that was congruent with their social roles and carefully designed their talk so that it was role-appropriate. It should

also be noted that *social role management* is a joint process by both parties, as evidenced by the duplicative organization 'father-son-uncle' in Brian's example. However, as Brian skilfully demonstrates, test-takers have the ability to utilize categories and categorization to their advantage in an interactionally relevant fashion. Exemplar 8 and Exemplar 9 feature a beginner-level test-taker Hans, who shows us that even low-proficiency L2 speakers have an awareness of *social role management* and can mobilize category knowledge to their benefit, despite limited linguistic resources.

Exemplar 8 Hans: 'you're a good kid'

```
1  H     Wo=wo=wo=wo   xiang:   ni    shi
         I  I  I  I    think    you   be
         'I=I=I=I thought: you were'

2  H  →  yi    ge   guai    hai (.) ZÌ (.) suoyi
         one   C    good    kid            therefore
         'a good ki(.)d(.) therefore'

3  H     (0.3)

4  H     Anm:(.)yexu (0.1) e:  (0.3)  YOU shihou (0.3) juede (0.5)
         PRT    perhaps    PRT        be  time         think
         'Hmm: (.)perhaps (0.1) Err: (0.3) There ARE times I think'

5  H     E:    gen    pengyou  he:    (0.1)  yi     liang  bei
         PRT   with   friend   drink         one    two    C
         'En:(.)you might have(0.1)one or two (0.3) drinks with friends'

6  H     (0.3)En (0.6) suoyi (0.2) bu  hui  you   wenti    dè=   buguo
              PRT      therefore   N   AUX  be    problem  NOM   but
         '(0.3)Hmm (0.6)therefore (0.2) it wouldn't be any problems=but'
```

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and H for Hans

Line 1 in Exemplar 8 follows Hans's story-telling of him receiving a call from the neighbour saying that lately there has been a lot of noise in his apartment. Hans makes multiple attempts at starting the turn but gets stuck at the subject *wo* 'I'. When he finally succeeds at the fourth attempt, he smoothly delivers his assessment of the interlocutor in line 2, calling him a *guai haizi* 'good/obedient

kid. Hans calling his friend's son *haizi* 'kid' positions the interlocutor at the lower end of a hierarchical relational pair compared to Hans, as Hans is the *grown-up* here while the interlocutor is the *kid* (Stokoe, 2012). Therefore, when a *haizi* misbehaves, a *grown-up* has sufficient moral grounds to criticize him, which is exactly what Hans's turn foreshadows.

Another interesting feature is the adjective *guai* that Hans predicates on *haizi*. *Guai* is a cultural-specific description that does not lend itself well to translation. Semantically it falls between 'being good', 'being nice', 'being obedient', 'do not cause trouble' and 'do not contradict'. *Guai* can only be used by a senior member on a junior member in Chinese society and the senior and junior members are related on an intimate personal level. For example, a younger person in Chinese society would not use *guai* as a quality expected of a person older than them. An older person would also not randomly call out people younger than them as being *guai* as there is no entitlement for such an expectation. What licenses Hans's usage of *guai* is the same duplicative father-son-uncle organization explicated before. Similar to what Brian did with his categorization work, Hans's word choice *guai* carries a strong moral obligation as there is a preference for a category like *haizi* to co-select with its bound predicates like *guai*. *Guai* as a predicate is a cultural expectation of the *haizi* category, which in this particular instance, manifests as 'being nice to *shushu*', 'do not contradict *shushu*' and certainly 'do not drink irresponsibly and cause trouble for *shushu*'. Therefore, Hans's attempt at characterization is purposeful as now the interlocutor's behaviour is not just unsociable but also flouts the moral obligations implicated in his category.

In terms of Hans's production of *guai haizi*, it is smoothly delivered in line 2, despite a micropause between *hai* and *zi*. This contrasts greatly with the following line 3-line 6 where Hans struggles to describe what the interlocutor was doing. Considering Hans's beginner-learner proficiency level, it is understandable that he finds it challenging to mobilize language in his recount of the interlocutor's irresponsible drinking episodes. However, Hans easily formats a predicate+category combination (*guai haizi*), which suggests that such a categorization, its bound predicate and inferences are not constructed on the spot. This combination is locally regulated, culturally pre-packaged, and contextually embedded in the very social role of Hans vis-à-vis the role of the interlocutor.

5.2.4.9 The power of categorization

Lastly, Exemplar 9 showcases the power of categorization in facilitating the fulfilment of the IC task when test-takers effectively orient to the categories and

their bound activities in the social roles. In the case of Hans's performance, the *guai haizi* categorization was not a one-off isolated incident as Hans also made it interactionally relevant in his closing sequence.

Exemplar 9 Hans: 'be a good kid'

```
1    I    Wo  hui=wo=>wo hui    zhuyi            de< (0.1)
          I   AUX I   I  AUX    pay attention    NOM
          'I will=I=>I will be more careful< (0.1)'

2    I    °Bu  hao  yisi°    (0.3) °Hai                    shushu°
           I'm sorry              (Hans's family name)     uncle
          '°I'm sorry° (0.3) °Uncle Hai°'

3         (1.7)

4    H →  Hao     a: (0.3) Guai:
          good    PRT      obedient
          'Ok: (0.3) Be good:'

5         (0.3)
6    I    °E (0.1) hao     de°
          PRT      good    NOM
          'Oh (0.1) Okey'

7         (0.6)

8    I    Xin   na    wo=°na     wo  yihou    hui    zhuyi           de°=
          Ok    then  I   then   I   future   AUX    pay attention   NOM
          'Okey, so I=°so I will be more careful in the future'

9    I    =na      xiexie  shushu
           then    thank   uncle
          '=so thank you uncle'
```

```
10        (1.1)

11   H    En
          PRT
          'Ok'

12        (0.6)

13   I    Hao      de
          Good     NOM
          'Okey'
```

Note. Pilot test_item 7, Chinese, video-chat, audio-recorded, role-play data, I for interlocutor and H for Hans

Exemplar 9 is towards the end of the interaction and starts off with the interlocutor promising not to misbehave in the future. After issuing a positive assessment *hao* 'good' in line 4, Hans recycles *guai* with emphasis and elongation, only this time as an imperative. *Guai* here takes on different functions as before: it is an assessment as in 'I underwrite what you just promised as *guai* behaviour'; it is also a veiled admonition as in 'be a *guai* kid and don't misbehave again'. In line 6 the interlocutor first issues a change-of-state token *E*, which is similar to 'oh' in English (Heritage, 1990). This token could have been occasioned by surprise at the warning tone in Hans's *guai*, or incredulity at Hans's highly idiomatic use of *guai* and his cultural knowledge of how Chinese people reprimand kids. Regardless of the mental state behind *E* 'oh', the interlocutor quickly proffers an agreement token and after a 0.6 gap, reformats his promise from line 1 in line 8 and thanks Hans for his assessment/admonition in line 9. The interlocutor's behaviour from line 6 to line 9 ratifies the legitimacy of Hans's use of *guai* and Hai's implicated moral rights. It would be highly problematic if the interlocutor rebuts by saying 'hang on, why should I be *guai* or be a *guai haizi*' or 'who do you think you are to talk to me in such a patronising manner'. This is synonymous with a mum saying 'why should I pick up my baby just because they are crying' in Sacks's example (Sacks, 1974). It is the inherent moral order residing in categories that makes the claims from category-bound predicates irrefutable. A similar example of the *haizi* category evocation plus a category-bound predicate is line 2 in Exemplar 5 where the test-taker categorizes the interlocutor as an 'understanding kid' and the analysis is very similar to the one of Hans.

In summary, Exemplar 4 to Exemplar 9 provide specific empirical evidence on how the indigenous rating category *social relations* can be theorized to

social role management and validated via test-taker discourse, using concepts from CA and MCA. DEs' focus on this rating category reinforces the argument that is no language use unbounded to social roles. Hence, the ability to enact appropriate social roles should be integrated into the IC construct. Brian's and Hans's performances indicate that this ability is largely unaccounted for in existing speaking/LC frameworks. From CA transcription we can observe numerous cases where Brian's and to a much larger extent Hans's language are linguistically inadequate, which include excessive long gaps, awkward turn designs, infelicitous lexical choices and non-L1 prosodic features[9]. However, inadequacies in speaking/LC do not obfuscate their ability in talking/IC, to undertake *social role management*, or to make their enacted social roles work for them interactionally. This mismatch between speaking competence/LC and talking competence/IC can go a long way towards strengthening the construct validity of IC assessment. The same principle applies to the other four rating categories, *disaffiliation control*, *affiliation promotion*, *morality* and *reasoning* (see Table 35), though more detailed analyses of these four categories are not provided here due to space restrictions. The CA-informed exemplars generated in the empirical validation can also serve as rater training materials to assist raters to connect descriptors in the rating scale with actual performance and to better understand the test construct.

5.2.5 A theorized IC rating scale

Section 5.2.4 shows how DEs' comments and their indigenous IC rating scale are validated in test-taker discourse. The process of using CA and MCA concepts to unpack DEs' comments in test-taker performance demonstrates that DEs' atheoretical comments and indigenous criteria can be theorized and interpreted from the CA and MCA perspective. The exemplars illustrated in Section 5.2.4 only represent a select number of CA- and MCA-analysed exemplars that study two produced. Based on more extensive analyses of DE comments, descriptors in the indigenous rating scale, and CA transcripts of test-taker performance, I transformed the DEs' indigenous scale in Section 5.2.3 into a theorized scale in this section, drawing on CA and MCA concepts. The process of theorization expands the applicability of the original indigenous rating scale, which is based on DEs' specific comments on specific test tasks from a specific cohort of test-takers. By theorizing through sociological theories, the new rating scale evokes

9 Some words that are pronounced with incorrect tones are marked with diacritics in the transcript.

concepts that have a wider range of applicability since the concepts are not dependent on specific assessment tasks or test-taker cohorts. This addresses the context-specific nature of existing IC assessment rating scales as the theorized rating scale is theoretically expandable enough to be applied to a range of assessment tasks and contexts. The new theorized scale still maintains the same five-step structure and includes five rating categories, each of which contains several sub-categories. The following sections explicate the theorized rating scale based on each of the five rating categories. A comparison between the theorized categories and the indigenous categories in Section 5.2.3 can highlight the differences between band descriptors derived from DEs' indigenous atheoretical descriptions and band descriptors informed by CA and MCA concepts (see Table 35 for a simplified presentation of the equivalent descriptors in two scales). The finalized theorized rating scale is presented in Appendix IV in both the English version and the Chinese version.

5.2.5.1 Theorized rating category: Disaffiliation control

The original *conflict management* rating category in Section 5.2.3.4 in the indigenous IC scale is theorized as *disaffiliation control*, which is more precisely defined through Sequential-Categorial Analysis. Based on the iterative analyses of the indigenous scale, DE comments and CA-transcribed test-taker performances, three sub-categories emerged in this rating category: 1) *Disaffiliative actions* (DAA), 2) *Linguistic disaffiliation control* (LDC), and 3) *Paralinguistic disaffiliation control* (PDC). The first sub-category, *disaffiliative actions*, discusses the use of disaffiliative actions at a more global level, such as the launch of complaining, criticizing, disagreeing, or refusing. These actions are disaffiliative and, if mishandled, can damage social solidarity (Steensig & Drew, 2008). Since the nine assessment tasks in the IC test all have disaffiliative scenarios and expect test-takers to launch some form of disaffiliative actions, sub-category 1, therefore assesses how the disaffiliative actions are designed by the test-takers, from a very indirect approach (band 5) that does not affect social solidarity to a very direct approach (band 1) that damages social relations.

As sub-category 1 is concerned with how disaffiliative actions are launched at a global level, sub-categories 2 and 3 focus on the linguistic and paralinguistic design of disaffiliative actions. Sub-category 2 *Linguistic disaffiliation control* is concerned with the linguistic devices test-takers use such as lexical choices, morphosyntactic features and dispreference formats. Sub-category 3 *Paralinguistic disaffiliation control* assesses paralinguistic devices such as phonetic and prosodic features as these features have been shown to match with

the linguistic design of the turns in managing disaffiliation (Couper-Kuhlen, 2012; Lindström & Sorjonen, 2013).

The rating scale for disaffiliation control is presented in Table 36, with each of the three sub-categories listed and their abbreviations (DAA, LDC, and PDC) included for clarity.

Table 36 Theorized IC rating scale for Disaffiliation control[10]

Band 5 Exemplary	• Disaffiliation is successfully and skillfully remediated, social solidarity unaffected. Disaffiliative actions are either unstated or approached exceedingly strategically in a way that does not cause affront (DAA). • Turn design conforms clearly with dispreferred formats, drawing on a wide range of lexical devices, morphosyntactic features and evidential markers (LDC). • Excellent control of phonetic and prosodic features in mitigating disaffiliative stances (PDC).
Band 4 Good	• Disaffiliation is well managed. Strong disaffiliative actions are covertly realized. Weak disaffiliative actions can be explicitly delivered but are admissible in the local context (DAA). • Both lexical and non-lexical devices are recruited to mitigate disaffiliation. Infelicitous choices are very occasional and not interactionally disruptive (LDC). • Phonetic and prosodic features are in tune with turn design in moderating disaffiliation (PDC).
Band 3 Average	• Disaffiliation is hearable though social solidarity is still intact. There is inconsistency in the speaker's management of disaffiliative activities (DAA). • Some conventionalized verbal resources are involved to soften disaffiliation to an acceptable level (LDC). • The phonetic and prosodic marking of turns has a limited impact on minimizing disaffiliation (PDC).
Band 2 Concerning	• A disaffiliative stance can be imputed to the speaker and interactional continuity can be disrupted. The use of disaffiliative actions and underuse of strategies can create or escalate tension (DAA). • There is recruitment of disaffiliative linguistic devices while the ones constituting a dispreferred format are inadequately represented (LDC). • The speaker's talk contains disaffiliative segmental and suprasegmental features (PDC).

10 From well-managed or no display of disaffiliation to bald-on-record display of disaffiliation.

Table 36 Continued

Band 1 Intervention needed	• Social solidarity is jeopardized due to the deployment of ostensible disaffiliative actions. The speaker's approach projects a clearly disaffiliative and uncooperative stance (DAA). • The design of turns lacks basic dis-preferenced features. Highly disaffiliative linguistic devices exist (LDC). • Turns are marked by pronounced disaffiliative segmental and suprasegmental features (PDC).

5.2.5.2 Theorized rating category: Affiliation promotion

The second rating category, *affiliation promotion*, is theorized from the indigenous IC category *solidarity promotion* in Section 5.2.3.5. As discussed in various places, *affiliation promotion* is different from *disaffiliation control* in that the former focuses on the actual promotion of social solidarity through the employment of affiliative actions, whereas the latter is concerned with how unavoidable disaffiliative actions can be managed to downgrade the destruction it has on social solidarity. Four sub-categories were identified under the affiliation promotion rating category: 1) *Affiliative actions* (AFA), 2) *Affiliative turn design* (ATD), 3) *Emotions and frames* (EAF), and 4) *Intersubjectivity* (ITS).

Similar to the sub-categories in *disaffiliation control*, the first category *Affiliative actions* in *affiliation promotion* focuses on the employment of social actions that are affiliative and supportive of social solidarity such as agreeing, praising, and apologizing. Sub-category 2 *Affiliative turn design* is more concerned with the actual lexical and non-lexical devices used in the affiliative actions. Sub-category 3 *Emotions and frames* foregrounds the demonstration of emotional engagement with the interlocutor (Haugh & Obana, 2015; Hayashi, 2013) and the sharing of their frames (Goffman, 1974; Kim, 2013). Lastly, sub-category 4 gives weight to the IC marker, *intersubjectivity*, which has been well researched in previous IC assessment research (Youn, 2015). Here, intersubjectivity is included under *affiliation promotion* because its pro-social nature is emphasized in the maintenance and promotion of social solidarity.

Table 37 presents the theorized IC rating category *Affiliation promotion* with abbreviations (AFA, ATD, EAF, and ITS) noted for each of the four sub-categories.

Table 37 Theorized IC rating scale for Affiliation promotion[11]

Band 5 Exemplary	• The speaker is maximally pro-social through successful launches of substantively affiliative actions (AFA). • An impressive array of lexical, morphosyntactic, phonetic and prosodic elements are skillfully mobilized in designing preferred actions (ATD). • A very high degree of empathetic understanding and sharing of frames is evoked, established, and maintained via the use of substantive forms of empathy display and frame identification (EAF). • The speaker is extremely keen and competent in pursuing, maintaining, and restoring intersubjectivity (ITS).
Band 4 Good	• Social solidarity is supported through the employment of commonly used affiliative actions (AFA). • The speaker can use (non)lexical devices to display affiliation. Phonetic and prosodic marking are in general in tune with turns (ATD). • There is sufficient recognition, validation and understanding of the interlocutor's emotions and frames (EAF). • Threats to mutual understanding are adequately addressed in the interaction (ITS).
Band 3 Average	• There is some use of affiliative actions to support the interlocutor's affective stance (AFA). • Turn design displays some conventionalized (non)verbal affiliation promotion (ATD). • Attempts are made to affiliate with the interlocutor's emotions and frames, though not always successful (EAF). • Intersubjectivity is sometimes checked and defended (ITS).
Band 2 Concerning	• Affiliative actions are underused. The interlocutor can feel somewhat unendorsed or unsupported (AFA). • (Non)verbal affiliative devices can be inadequate or mis-designed (ATD). • Little display of empathetic understanding of the interlocutor's feelings and experiences (EAF). • The speaker's understanding is prioritized while common understanding is largely unaddressed (ITS).
Band 1 Intervention needed	• Little to no indexation of affiliation. The interlocutor's affective needs are disregarded (AFA). • Preferenced (non)lexical affiliative features[12] are rare to non-existent. Little to no prosodic matching or upgrading for affiliative work (ATD). • Noticeable misalignment from and disaffiliation with the interlocutor's emotions or frames (EAF). • The speaker overtly privileges their understanding and is disinterested in achieving intersubjectivity (ITS).

11 From overt display of affiliation to no show of affiliation.

12 Preferred, preferenced and affiliative features are supportive of social solidarity in both 1st pp or 2nd pp positions.

5.2.5.3 Theorized rating category: Morality

The rating category *morality* is derived from the category *personal qualities* in Section 5.2.3.6 in the original indigenous IC rating scale. As illustrated in Exemplar 3 in Section 5.2.4.5, the category *personal qualities* is expanded and further specified from an ethnomethodological perspective in the sense of moral order being constitutive of social order.

The definition of morality in this book draws on the distinction between proto-morality and cultural-specific moralities in previous literature (Bergmann, 1998; Turnbull & Carpendale, 2001; Turowetz & Maynard, 2010) but expands it and builds a three-tier conceptualization of moral order: 1) *Endemic moral order* (EMO), 2) *Universal moral order* (UMO), and 3) *Context-specific moral order* (CMO). At the basic level, which is the sub-category 1 *Endemic moral order*, morality is endemic to the discourse and refers to the cooperative, reciprocal principle of interaction between interactants, which makes interactional order possible at a local level. This is similar to the concepts of *protomorality*, *moral order 1* and *morality of interaction* as discussed in existing literature in Section 2.4.4.

Sub-category 2 *Universal moral order* defines a moral order that is across contexts and cultures, and that is built on the local endemic moral order and allows interactants to infer moral qualities of their interlocutors. This is similar to the morality *in* interaction, defined in Turowetz and Maynard (2010). The moral qualities inferred from the moral order at this level are not context-dependent, as they are qualities such as respect, dignity, courage and composure, as listed in Hymes (1972) and discussed in Section 2.4.3. These are qualities that are universally mandated in all interactional contexts, and once breached, can attract moral sanctions and condemnations regardless of the specific interactional contexts.

The last sub-category, sub-category 3 *Context-specific moral order*, is concerned with the moral order at a context-specific level. This moral order is in operation when we infer context-specific or cultural-specific qualities from our interactants. They are the qualities such as humbleness, modesty and collectivism that were attended to by the DEs, who were from a specific linguistic and sociocultural background. Such context-specific moral values are contingent on the interactants' shared life experiences and the interactional contexts, which means that the moral values in sub-category 3 can be realized and interpreted differently when the context changes.

How the three-tier conceptualization of morality is realized at five different band levels is illustrated in Table 38. Sub-categories in the morality rating scale are noted with their abbreviations (EMO, UMO, and CMO).

Table 38 Theorized IC rating scale for Morality

Band 5 Exemplary	• The speaker's manner of interaction is maximally cooperative, fully sustaining the endemic moral order of interaction (EMO). • The interaction clearly indexes universally preferred moral qualities in the speaker (UMO). • The speaker's moral membership is fully established through the demonstration of moral conduct specific to the standards in their community (CMO).
Band 4 Good	• The moral order of interaction is well sustained through a focus on cooperation instead of constraint (EMO). • The speaker's conduct can allow for the ascription of some universally preferred moral qualities in the speaker (UMO). • The speaker's conduct showcases good understanding of moral standards known to members of the shared community (CMO).
Band 3 Average	• Interactional order is in general maintained though some conducts are not fully morally accountable or accounted for (EMO). • There is no strong indexation of either universally preferred or dispreferred moral qualities (UMO). • Community-specific moral conduct is lacking, though the speaker's moral membership is still accepted (CMO).
Band 2 Concerning	• Interactional order is threatened by noticeable moral breaches (EMO). • Some universally dispreferred moral qualities can be ascribed to speaker (UMO). • The speaker's conduct is recognizably morally questionable to members in the community (CMO).
Band 1 Intervention needed	• Moral order is disrupted by a constraining interactional manner and morally unaccountable conduct (EMO). • The speaker deploys universally morally sanctionable actions that challenge their status as a moral being (UMO). • The speaker faces ostracization due to the display of morally unjustifiable conduct in the community (CMO).

5.2.5.4 Theorized rating category: Reasoning

The theorized rating category *reasoning* is built on the indigenous rating category *reasoning skills* in Section 5.2.3.7 but draws on concepts from CA and ethnomethodology (Garfinkel,1967; Heritage, 1988). This category is

further specified into three sub-categories: 1) *Interactional remedies* (IAR), 2) *Interactional accounts* (IAA), and 3) *Overall structural organizations* (OSO).

Sub-category 1 *Interactional remedies* is concerned with how interactional troubles are remedied at both the action and project levels. Actions here refer to the specific social actions and the sequences during which they are realized. Projects, on the other hand, are sequences of sequences and can include multiple social actions (Robinson, 2013). Remedies here are defined as solutions that speakers proffer to address interactional troubles, whether within the turns in a social action or within a larger interactional project. The differentiation between action and project here mirrors the distinction between the two levels of reasoning discussed in Heritage (1988), although Heritage did not further define the types of reasoning at the two levels whereas in the definition of *reasoning* in this book, reasoning sub-category 1 focuses specifically on the remedies speakers can provide to address interactional troubles.

Similar to the scope in sub-category 1, sub-category 2 *Interactional accounts* also pertains to both the action and the project levels but focuses more specifically on the accounts the speaker provides when accounts are due. Accounts are defined as the explanations that speakers proffer when the context or the interlocutor demands explanations for the speaker's action.

The last sub-category, sub-category 3 *Overall structural organizations*, extends on what DEs called *the structure of talk* in the indigenous rating scale in Section 5.2.3.7. The concept *overall structural organization* (Robinson, 2013) in CA captures DEs' descriptions of the structure of a person's speech. Overall structural organization is concerned with how multiple social actions are organized since a stretch of talk rarely contains only one action. In response to an interlocutor's invitation to a dinner party, a speaker may first thank the interlocutor for the kind invitation, apologize for not being able to attend, explain the reasons behind it and wish the interlocutor a successful party or promise to attend when the next opportunity presents itself. It is clear this talk contains multiple social actions such as thanking, apologizing, explaining, and offering wishes/promises. Overall structural organization examines the way different social actions are strung together. In other words, it looks at the organization of sequences of sequences (Schegloff, 2007b).

How the three sub-categories are defined at each of the five bands is presented in Table 39 with the sub-category abbreviations included (IAR, IAA, and OSO).

Table 39 Theorized IC rating scale for Reasoning

Band 5 Exemplary	• Extremely efficacious, multi-angled and context-fitting remedies are provided to address interactional troubles at both the action and project levels (IAR). • The speaker provides exceedingly well-reasoned and contextually defensible accounts at both the action and project levels (IAA). • The entire occasion of interaction is exceptionally well organized at both the sequential and sequence levels, providing maximal progressivity and projectability (OSO).
Band 4 Good	• The speaker proffers quality and applicable remedies to troubles, though slightly more substantiation of the remedies would be ideal (IAR). • The accounts proffered conform with the shared reasoning in a normative social life (IAA). • Interaction is structured in a recognizable manner around a beginning, different units of interaction and a closing (OSO).
Band 3 Average	• The remedies offered have limited impact on addressing the troubles in the interactional context (IAR). • Accounts are provided but are not the most readily acceptable ones or are in need of development (IAA). • The overall structural organization is in place, though the design of some interactional units can be improved (OSO).
Band 2 Concerning	• The remedies are simplistic or ineffective. Interactional troubles are largely unaddressed (IAR). • Noticeable limitations exist in the accounts supplied, dis-conforming with the shared reasoning among interactants (IAA). • There are noticeable issues with the organization of sequences and the organization of sequences of sequences (OSO).
Band 1 Intervention needed	• Either no remedies are provided, or the remedies provided are problematic or context-misfitting (IAR). • Either no accounts are proffered, or the accounts supplied are poorly reasoned or norm-defying (IAA). • The organization of interaction is weak, actions are unrecognizable, and progressivity is stalled (OSO).

5.2.5.5 Theorized rating category: Social role management

The last rating category, *Social role management*, is based on the indigenous rating category *Social relations* in Section 5.2.3.8 and is informed primarily by MCA concepts and the MCA part of the Sequential-Categorial Analysis of the exemplars in Section 5.2.4. This category is comprised of three sub-categories: 1)

Categories and predicates (CAP), 2) *Broader membership apparatuses* (BMA), and 3) *Social role competition* (SRC).

The first sub-category, *Categories and predicates*, focuses on test-takers' ability to match categories with activities/predicates for not only the categories in the roles they undertake but also in the roles their interlocutors assume. This ability is the core of MCA in that it elicits test-takers' member knowledge of what members in particular categories/social roles are supposed to do and act in particular contexts (Sacks, 1974; Stokoe, 2012). Predicates can also go beyond activities commonsensically associated with certain categories and include duties, responsibilities and obligations (Jayyusi, 1984; Payne, 1976; Watson, 1978).

The second sub-category, *Broader membership apparatuses*, taps test-takers' knowledge of the broader social relation devices that regulate the categorial aspect of interaction. Here I define broader membership apparatuses as including standardized relational pairs, membership categorization devices and duplicative organizations (Stokoe, 2012). Knowledge of these apparatuses goes beyond what a test-taker knows about specific categories and their category-bound predicates. The focus of broader membership apparatuses is test-takers' knowledge of how categories interact and are positioned within broader social relations.

Lastly, *Social role competition* is a sub-category that was theorized from DEs' comments (see the sub-category 4, *mediating multiple social roles* in the original indigenous rating scale in Section 5.2.3.8). *Social role competition* is a phenomenon that has not been discussed extensively in the IC literature but is one that captured DEs' attention. Here I term role competition as the tension between the roles that the speaker or the interlocutor needs to assume. For example, a test-taker can be both 'a landlord' and 'the tenant's father's best friend'. Test-takers cannot only address one of the roles here as both roles evoke specific category-bound predicates that can work to test-takers' advantage but also simultaneously circumscribe test-takers' conduct. The ability to mediate the multiple and sometimes competing categories in test-takers' social roles and their interlocutors' social roles is crucial for successful social role management. Table 40 presents the IC rating scale for *Social role management*, including all three sub-categories, their abbreviations (CAP, BMA, SRC), and their accompanying descriptors.

Table 40 Theorized IC rating category for Social role management

Band 5 Exemplary	• The speaker has outstanding competence in enacting and orienting to social roles that are highly congruent with their and the interlocutors' categories, matching their conduct with their respective category-bound predicates and relative hierarchical positioning (CAP). • There is an excellent application of the standardized relational pairs, membership categorization devices and duplicative organizations to which the speaker and interlocutor belong (BMA). • Highly skilled mediation and prioritization of the speaker's and the interlocutors' roles are demonstrated (SRC).
Band 4 Good	• The speaker demonstrates the ability to enact and orient to relevant social roles. The matching of category-bound predicates is overall felicitous (CAP). • The speakers can utilize some broader membership apparatuses (BMA) to make their conduct recognizable (BMA). • Role competition is balanced to achieve successful interaction (SRC).
Band 3 Average	• The expected roles are in general enacted and oriented to. Some predicates can be over-realized or under-realized (CAP). • The application of BMA is limited but no misuse exists (BMA). • The speaker's and the interlocutors' primary roles are oriented to, but other roles are insufficiently addressed (SRC).
Band 2 Concerning	• Social role management is insufficient. Categories are not well matched with predicates (CAP). • There are incidents of misapplications of BMA, suggesting a lack of knowledge of what context-fitting BMA to draw on (BMA). • Primary roles are not adequately attended to (SRC).
Band 1 Intervention needed	• Normatively expectable categories are not attended to and there is grave misunderstanding regarding category-bound predicates (CAP). • The speaker neglects context-relevant BMA or seriously mismanaged BMA, disrupting the interaction (BMA). • There is mis-prioritization of roles and role competition is overlooked (SRC).

5.2.6 A unified model of IC

The theorized IC rating scale in Section 5.2.5 presents a scale that is rooted in DEs' indigenous criteria but at the same time, is theoretically robust since it is undergirded by sociological theories. It has the potential to be applied to a range of assessment settings due to its theoretical underpinnings, and also be specified to localized, context-specific test tasks. The IC test construct the scale embodies is presented in Figure 19.

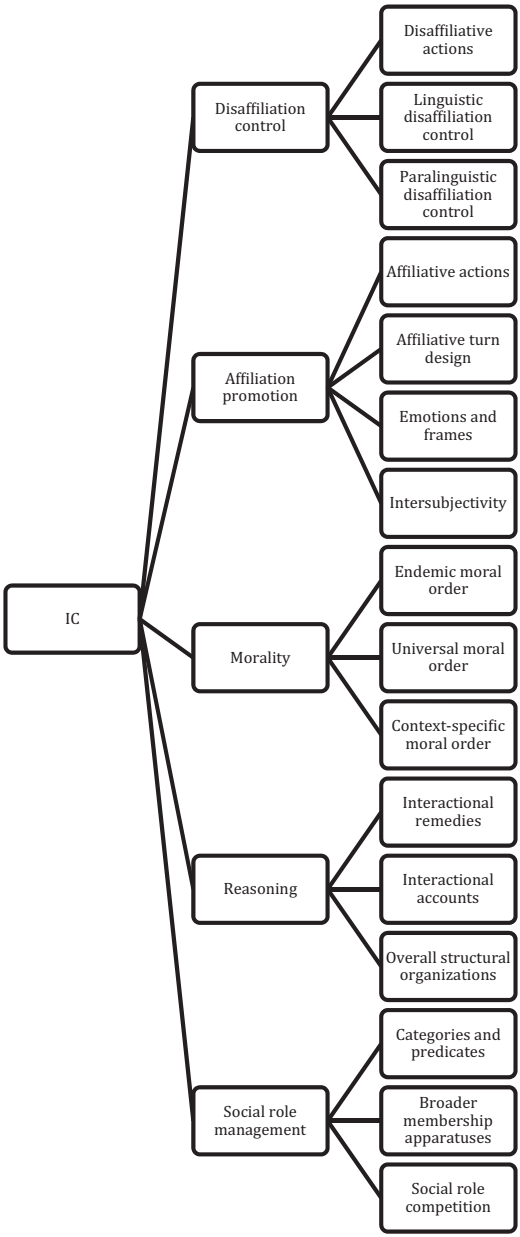


Figure 19 A schematic representation of the IC test construct

If we revisit the Hymes-ean definition of communicative competence and the three artistic proofs in Aristotelian rhetoric, as discussed in Section 2.4 in the literature review, we can see that the five rating categories encompass the components in both models. The IC construct arrived at in this book covers the emotional, logical, moral and personal aspects of interaction that were integral to Hymes's conceptualization of communicative competence. At the same time, the five categories correspond to the emotional, logical, and moral dimensions of Aristotle's model of effective persuasion, mirroring the expanded definitions of pathos, ethos, and logos. More specifically, pathos refers to emotional engagement between interactants, ethos the projection of moral qualities of the *interaction* and of *interactants*, and logos, the logic *in* and *of* interaction. The semblance between the IC construct in this book and the two interpersonal communication models in sociology and philosophy attests to the theoretical robustness of the five rating categories. Whether we look at interaction from a philosophical, sociological, sociolinguistic or ethnomethodological perspective, we are still examining the same phenomenon: inter-action. It is therefore expected that there is some overlap between different models of interpersonal communication, and the fact that the IC construct captured in the theorized rating scale overlaps with other holistic models of interaction is evidence that the IC construct in this book is comprehensive in its coverage of aspects of interaction.

Figure 20 is a schematic representation of how IC can be inferred from test-taker performance via the five IC rating categories. The dashed line represents the inference of the unobservable IC from the observable scores in the five rating categories. Both *disaffiliation control* and *affiliation promotion* fall under *pathos* as both of them address emotional engagement with the interlocutor, albeit from opposite angles. *Morality* and *reasoning* correspond to *ethos* and *logos*. At the top of the five rating categories is *social role management*, which is a higher-order IC indicator as it encompasses an understanding of the categorial aspect of interaction. In the theorized IC scale there are descriptors pertaining to the mechanic, sequential properties of interaction (May et al., 2020) but their functions to IC are reconceptualized as facilitating and mediating the realization of the other aspects of interaction, which are more salient to real-world everyday-life interactants. Assessing sequential markers of IC alone, such as turn-taking and back-channelling[13], can obscure the social functions sequence fulfils in interaction. Therefore, the term sequential is placed in the middle of the circle

13 The IC tree in Galaczi and Taylor (2018) offers comprehensive coverage of many of the frequently researched sequential IC markers.

to emphasize the enabling nature of sequence for the display of the emotional, logical, moral, and categorial dimensions of interaction.

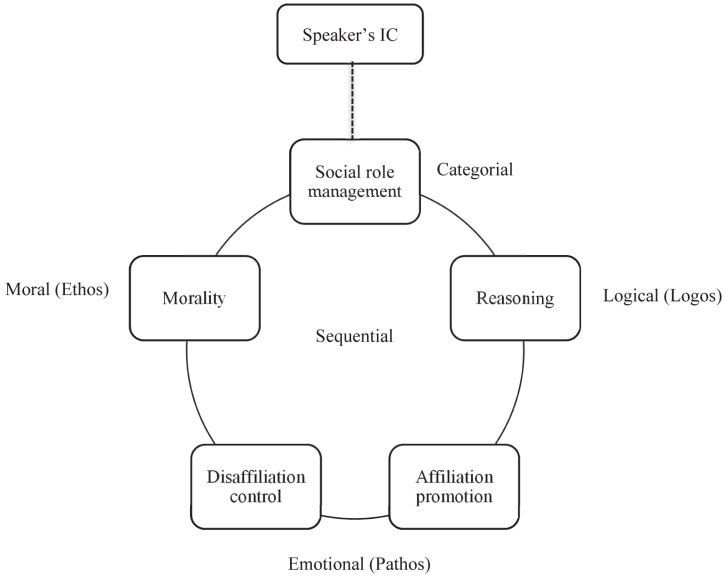


Figure 20 Theoretical import of the five IC categories

Chapter 6 Study three: The IC test and accompanying questionnaires

Chapter 6 reports the research design and findings of study three, which is the main testing study. In this study, the IC test developed in Chapter 4 was put to use with a large cohort of test-takers and the rating materials developed in Chapter 5 were used by raters to assess test-taker performance. The rationale for conducting study three and addressing the research questions of study three are detailed in Section 3.2.3.

105 test-takers participated in study three; they completed the IC test, a self-assessment questionnaire of IC, and nominated an L1-Chinese peer to rate their IC in a peer-assessment questionnaire. Test-takers' performances on the nine-item IC test were rated by two raters in a fully-crossed design. The rating results, analysed through many-facet Rasch, were satisfactory in terms of candidate measurement, rater measurement, rating category measurement, item measurement and rating scale functioning. Rasch residuals principal components analysis (PCA) demonstrated the unidimensionality of the data structure, evidencing that the five rating categories in the theorized IC scale, (1) disaffiliative control, (2) affiliation promotion, (3) morality, (4) reasoning and (5) social role management, worked together in their measurement of a unidimensional IC construct.

Having analysed test-taker test performance using Rasch, I correlated test-taker test performance with results of their self-assessment and peer-assessment IC questionnaires. There was a significant correlation between test-taker IC test performance and their self-assessment of their IC in the wild (Hellermann et al., 2019), though the correlation between test performance and test-taker peers' assessment was not significant. The significant self-assessment results take precedence over the non-significant peer-assessment ones due to the oftentimes unreliable nature of peer assessment. This shows that the IC test can reasonably predict test-takers' IC in real-life, non-testing settings.

Finally, test-takers' attitudes towards the test and perception of the extrapolative strength of the test were also ascertained in the test-taker self-assessment questionnaire. Findings show that test-takers welcomed the format of this CMC-based IC test and believed the test can offer a good prediction of their IC in real-world scenarios.

6.1 Methodology

Section 6.1 is the methodology section that explains the participants, instruments, procedures, and data analysis of study three.

6.1.1 Participants

6.1.1.1 Main testing test-takers

Different from the relatively smaller number (N=22) of test-takers in the pilot test in Chapter 5 (see Section 5.1.1.1 for details), the main testing session in study three required a larger number of test-takers to better investigate the psychometric qualities of the IC test. 105 test-takers from 26 countries participated in the main testing session. Their home countries are listed in Table 41 ranked by the number of test-takers from that particular country. As I was based in Australia when conducting study three, it was easier to advertise the test to local Australian L2-Chinese speakers, hence the larger group of test-takers from Australia. The 15 test-takers from China were L1 speakers of Chinese and were recruited for comparison. The recruitment of test-takers from a wide range of home countries was facilitated by the CMC nature of the IC test, which allowed test-takers to participate in the test online without having to travel to a testing site. It is also worth noting that more than half of the countries (N= 14) where the test-takers came from are countries that receive ODA (Official Development Assistance) on the Development Assistance Committee (DAC) list produced by the Organization for Economic Co-operation and Development (OECD) (OECD, 2021). It shows that the accessibility of CMC assessment has the capacity to reach economically disadvantaged test-takers and test users, improving the fairness and affordability of language assessment. Insight from under-represented countries and test-taker groups can also democratize and enrich the research basis of language assessment.

Table 41 Main testing test-takers' home countries

Main testing test-takers' home countries	N	%
Australia	20	19.05 %
China	15	14.29 %
Vietnam	11	10.48 %
USA	9	8.57 %
Italy	5	4.76 %
Pakistan	5	4.76 %

Table 41 Continued

Main testing test-takers' home countries	N	%
Thailand	5	4.76 %
Austria	4	3.81 %
UK	4	3.81 %
Israel	4	3.81 %
South Korea	4	3.81 %
Kazakhstan	3	2.86 %
Belgium	2	1.90 %
Spain	2	1.90 %
Myanmar	1	0.95 %
Cambodia	1	0.95 %
France	1	0.95 %
Germany	1	0.95 %
Indonesia	1	0.95 %
Iran	1	0.95 %
Morocco	1	0.95 %
Nigeria	1	0.95 %
Russia	1	0.95 %
Tadzhikistan	1	0.95 %
Ukraine	1	0.95 %
Uzbekistan	1	0.95 %
Total	**105**	**100.00 %**

The wide variety of test-taker home countries also ensured the test-takers were of various language backgrounds. The first languages represented in this 105-test-taker dataset are listed in Table 42. Seven test-takers grew up bilingually and both their languages are listed. The recruitment of test-takers with a wide array of L1 backgrounds minimizes the effect of L1 on test results, boosting the generalizability of the research findings.

Table 42 Main testing test-takers' first languages

Main testing test-takers' L1s	N	%
Australian English	20	19.05 %
Chinese	15	14.29 %
Vietnamese	11	10.48 %

(*Continued*)

Table 42 Continued

Main testing test-takers' L1s	N	%
American English	9	8.57 %
German	5	4.76 %
Thai	5	4.76 %
Urdu	5	4.76 %
British English	4	3.81 %
Italian	4	3.81 %
Korean	4	3.81 %
Russian	4	3.81 %
French	2	1.90 %
Spanish	2	1.90 %
Arabic	1	0.95 %
Burmese	1	0.95 %
Farsi	1	0.95 %
Hebrew	1	0.95 %
Indonesian	1	0.95 %
Khmer	1	0.95 %
Nigerian English	1	0.95 %
Ukrainian	1	0.95 %
Hebrew/English	2	1.90 %
French/English	1	0.95 %
Hebrew/Russian	1	0.95 %
Italian/English	1	0.95 %
Kazakh/Russia	1	0.95 %
Russian/Farsi	1	0.95 %
Total	**105**	**100.00 %**

In terms of other demographic information, Table 43 and Table 44 present the distribution of test-takers' genders and ages. Previous testing research tends to only recruit tertiary students within a narrower age range as their participants. The privileging of tertiary students as participants in applied linguistics research, due to the ease of access, can however underrepresent the diverse language users outside higher education settings and limit the generalizability of research findings. The convenience of the CMC test delivery format of this IC test allowed test-takers from a wide range of age groups and backgrounds to participate. The test-takers in study three ranged from tertiary students studying in China or

outside China, naturalistic L2-Chinese speakers who picked up the language from their friends or partners, to early-career and senior working professionals who used L2-Chinese for business purposes. CMC has allowed applied linguistics research to reach a wider group of interested participants, which helps to democratize research in the field.

Table 43 Main testing test-takers' genders

Main testing test-takers' gender	N	%
Male	60	57.14 %
Female	45	42.86 %
Total	**105**	**100.00 %**

Table 44 Main testing test-takers' age groups

Main testing test-takers' age groups	N	%
Above 40	6	5.71 %
Between 36 to 40	5	4.76 %
Between 31 to 35	14	13.33 %
Between 26 to 30	32	30.48 %
Between 20 to 25	45	42.86 %
Below 20	3	2.86 %
Total	**105**	**100.00 %**

6.1.1.2 Main testing test-taker peers

The 105 test-takers in this test were encouraged to nominate an L1-Chinese friend of theirs to complete a peer-assessment questionnaire. The purpose of administering a peer-assessment questionnaire is to elicit information regarding how test-takers interacted *in the wild* outside testing settings. The selection criteria for peers were that they needed to be L1 speakers of Chinese and had frequent interaction with the test-takers in Chinese. Out of 105 test-takers, 73 test-takers were able to refer L1-speaker peers who then completed the peer-assessment questionnaire. Some L2-Chinese test-takers were unable to nominate L1-Chinese peers because they either claimed to not have many L1-Chinese friends, not speak to them in Chinese often or simply not feel comfortable asking their L1-Chinese peers to assess their IC.

6.1.1.3 Main testing IC test raters

Two raters were recruited to rate test-takers' performance on the IC test using the IC rating scale. Both raters were L1-Chinese speakers and did not possess any training or experience in Chinese language teaching, CA or speaking assessment, making them linguistically naïve raters like the ones employed to rate the performances of the 22 pilot test-takers in Section 5.1.1.2. The rationale for selecting raters with no applied linguistics or language testing experience is that I hypothesized that raters who were not linguistically naïve were more likely to be influenced by test-takers' LC in their judgement of IC, given how deeply entrenched the psycholinguistic-individualist model of LC is in language teaching and testing in general. Choosing LC-naïve raters and providing them with carefully designed rater training materials on IC (see Section 6.1.2.4 for details) can maximize the chances of the raters focusing on IC features in their rating.

6.1.2 Instruments

6.1.2.1 The IC test

The IC test administered to all 105 test-takers is the test developed in Chapter 4. The specifications of the test can be found in Table 12 in Section 4.2.2. A summary of the test structure is in Section 4.2.5. A detailed description of the test construct is summarized in Figure 19 and Figure 20 in Section 5.2.6. The rating scale used for this test is presented in detail in Section 5.2.5 and the operational version of the scale in both English and Chinese is included in Appendix IV. A brief recapitulation of the items in the test is offered here for ease of understanding.

The IC test includes nine items of three task delivery methods, starting with three 1st pair-part voice messaging items, then three 2nd pair-part voice messaging items, and lastly three live video chat items. The three items within each task method group vary in the language use domain (study, work, everyday life) and the power variable (the test-taker having more power, equal power or less power, compared to the interlocutor). The order of the three items within each task method group is of increasing item difficulty based on Rasch findings from the pilot study in Section 5.2.1. The order of the three task methods follows the increase of interactiveness of the methods. 1st pair-part items only require test-takers to initiate interaction with the least amount of interlocutor interaction involved. 2nd pair-part items expect test-takers to supply responses to 1st pair-part speech from the interlocutor, which involves a higher degree of interactiveness. The last three items, video chat items, tap test-takers' real-time,

online IC as they need to interact with a live interlocutor synchronously. Hence, the last three items are the most interactive ones of all nine items.

The task scenario for each of the nine items is delivered via a short video consisting of three to four still images and an audio track narrating the scenario in Chinese. The images visually represent the task scenarios in a cartoon format (see Figures 16–18 for examples) so as to ensure lower-proficiency test-takers can fully comprehend the task scenarios. Video delivery of task scenarios also makes the IC test more engaging for test-takers. Keywords in the task scenarios are featured in both Chinese and English in the video images to aid comprehension, as Figure 16 to Figure 18 show. The audio narrating the task prompts was recorded by an L1-Chinese speaker trained in professional broadcasting in a professional recording studio. The speech rate was controlled at approximately 2.2 Chinese characters per second, which is lower than the normal Chinese speech rate (4.3 characters per second, see Chan & Lee, 2005). The rationale for lowering the speech rate was arrived at in consultation with Chinese teachers to ensure that listening comprehension of the task scenarios does not become construct irrelevant variance in the measurement of test-taker IC. The recruitment of a professional speaker, the utilization of a professional recording studio and the control of speech features ensure the test recordings are accessible to L2-Chinese speakers of differing proficiency levels. The still images, Chinese voice-over, and the English translation of the Chinese voice-over for each of the nine task scenarios can be found in Appendix III.

6.1.2.2 Test-taker background questionnaires

Before the test-takers in study three commenced the IC test, they completed a background questionnaire with the researcher. The background questionnaire asked test-takers about their demographic information such as age, first languages, country of origin and length of residence in China, which are presented from Table 41 to Table 44. The demographic information was collected to understand factors that can influence test-takers' IC.

6.1.2.3 Self and peer-assessment questionnaires

After test-takers completed the IC test, they were asked to finish a self-assessment questionnaire. The questionnaire has two sections. Section A consists of 11 questions that tap two separate constructs. The first construct is concerned with test-takers' perception of the IC test in terms of how much their performance on the test mirrors their real-world behaviour. Six items measure this construct at different levels of the construct, which means the items were designed to elicit

different degrees of endorsement from the respondents but at the same time measure the same construct (McNamara et al., 2019). This construct is related to the extrapolation inference in the validity argument.

The other construct is test-takers' attitudes towards the IC test in relation to the use of the test for teaching, learning, testing and other commercial purposes. Five items were designed to assess this construct. This construct is not directly related to any of the inferences in the validity argument as no inference examines stakeholder's attitudes towards and uptake of a test. This will be discussed separately in Chapter 8 outside the validity framework. The items for both constructs are presented in six-point Likert scales from *strongly disagree*, *disagree*, *slightly disagree*, *slightly agree*, *agree* to *strongly agree*.

Section B in the self-assessment questionnaire consists of 35 Likert-scale items designed to tap the five IC rating categories in the IC scale. Instead of assessing test-taker IC based on test performance in the IC test alone, I created these 35 questionnaire items to measure test-takers' self-assessment of their IC *in the wild*, which is non-testing settings in real life. I wrote the 35 items based on the most frequently mentioned codes in the IC categories in DEs' indigenous criteria in Section 5.2.2. Out of the 35 items, eight measure the rating category *disaffiliation control*, six for *affiliation promotion*, eight for *morality*, six for *reasoning* and seven for *social role management*., the 35 items similarly measure the five IC categories at different levels of the construct (McNamara et al., 2019). Each item is presented with a six-point Likert scale as in Section A but test-takers are also given the option of selecting 'unable to assess' for any of the items. Details regarding the questions for each IC category are presented in Section 6.2.3 where the questionnaire results are presented.

As mentioned in Section 6.1.1.2, each test-taker was invited to nominate an L1-Chinese peer to complete a peer-assessment questionnaire which assesses test-takers' IC in everyday life from a peer, non-testing perspective. The peer-assessment questionnaire contains the same 35 items as in Section B in the test-taker self-assessment questionnaire. The only difference is the propositions in the 35 items were phrased from the perspective of L1-Chinese peers as in 'I think my friend can…' instead of from the test-taker perspective in the self-assessment questionnaire.

Though the items in both self and peer-assessment questionnaires target specific theoretical constructs of IC, the items were written in plain language to ensure the questionnaires are accessible to test-takers and test-taker peers who are not familiar with the theories and concepts of IC. Both questionnaires are included in Appendix V and VI in English and Chinese.

6.1.2.4 Rater training materials

In terms of rater training, three types of training materials were developed from the pilot test-taker performances, on top of the theorized IC rating scale from Section 5.2.5 and in Appendix IV. Details of the three types of rater training materials and how they were used for rater training are provided in the rater training section in Procedures in Section 6.1.3.2.

6.1.3 Procedures

6.1.3.1 Administering the IC test and questionnaires

To recruit test-takers for the main testing session, I first posted advertisements regarding the IC test on online platforms. The uptake from L2-Chinese speakers was enthusiastic and the test-takers who participated in the test further promoted the test in their local networks and on their local online platforms. The interest from potential test-takers around the world was to such an extent that allowed me to use HSK and other demographic information as screening criteria to ensure a more heterogenous test-taker sample. The HSK criterion used was that there should be approximately 30 HSK 3 or below test-takers, 30 HSK 4–5 test-takers, 30 HSK 6 and above test-takers, and 15 L1 test-takers. Prospective test-takers needed to confirm their HSK levels to be able to proceed to take the IC test. There was one test-taker who had never sat any HSK or formal Chinese proficiency tests but had lived and worked in China for more than 10 years. He was a self-taught L2-Chinese with a jagged profile of being very strong in speaking and listening but limited in reading and writing. He was included in the 105 test-taker group to diversify the test-taker profiles. Compared to around 60 above HSK3 test-takers, only 30 HSK3 and below test-takers were chosen. The rationale for choosing relatively fewer test-takers with limited proficiency was because of the challenging nature of the IC test in this book which requires test-takers to have a relatively high level of linguistic competence (LC) to be able to interact in the test tasks. Other demographic information such as country of origin and gender was also taken into consideration to make the test-taker sample in the main testing study more representative of the L2-Chinese speakers *in the wild*.

Once the test-takers were selected, they were invited to schedule a time for a one-on-one testing session online via video chat. Prior to the testing session, I sought consent from the test-takers to gain their permission to have their test performances video-recorded and used for research and teaching purposes. After gaining consent, the test-taker completed the background questionnaire

(see Section 6.1.2.2 for details of the questionnaire). I then started administering the test by sending the test prompt for each of the nine items. After viewing the prompt videos, test-takers were given up to 60 seconds to prepare their answers. Once the time was up or when the test-taker indicated they were ready, they started sending voice messages for the first six items. In order to bolster task authenticity, there was no limit on how many voice messages each test-taker could send as in real life an interactant can send as many voice messages as one prefers to either initiate a conversation or respond to voice messages from others. For the video-chat items, once the preparation time was over, I would hang up the video chat and either initiate a new video chat or wait for the test-takers to initiate the video chat, depending on who was expected to make the video call in the task scenarios. This set-up, again, was meant to strengthen the authenticity of the tasks by imitating authentic video calls as required in the last three video chat test tasks.

After the test-takers completed the IC test, they proceeded to complete the self-assessment questionnaire. Once the self-assessment questionnaire was finished, the testing session for the test-taker was over. The test-takers were then asked if they could identify an L1-Chinese peer to complete the peer-assessment questionnaire. For the test-takers that could nominate L1-Chinese peers, they were given the peer-assessment questionnaires to pass on to their L1-Chinese peers, who subsequently filled out the peer-assessment questionnaire and passed it on to the researcher.

6.1.3.2 Training raters

As briefly mentioned in Section 6.1.2.4, I developed three types of rater training materials for the three steps in the training of naïve raters to rate IC performances.

Step one: Using **CA/MCA exemplars** to understand the rating categories. Six CA-informed exemplars (see Section 5.2.4) were assembled in an easy-to-follow manner to assist raters with comprehending the five IC rating categories and the descriptors in the rating scale. Raters were first presented with CA-informed test-taker exemplars and the Chinese version of the IC rating scale. I explained the rating scale and answered questions raters had regarding the content and wording in the scale descriptors. Then I played the audio for the exemplars and walked raters through the exemplars. I briefly explained the CA annotation system to the raters to facilitate their understanding of the interactional features worth noticing in the exemplars. As discussed in Section 5.2.4, the exemplars served as a crucial linkage between the rating scale and test-taker performance since the exemplars illustrate how the theory-laden descriptors in the rating

scale in Section 5.2.5 and Appendix IV can be reflected and located in actual test-taker interactional data. After the presentation of exemplars for the five rating categories, raters developed a preliminary understanding of how the five IC categories were defined and how their various features could be identified in test-taker discourse.

Step two: Using **ranking exercises** to develop sensitivities to differing levels of performance. The second step in rater training was the presentation and trial rating of sample performances on specific items for specific IC categories. Raters were sent five sets of eight sample performances, totalling 40 samples. For each set of samples, they needed to rate the eight samples in that set on only one IC category. The eight samples were deliberately chosen to be at differing performance levels of the target IC category. The purpose of this step was to train raters to notice differences in test-taker performances using the IC scale. The rationale for only asking raters to focus on one IC category per set and for only including performances on the same item in each set was to not overload raters cognitively when they were still familiarizing themselves with the steps in the rating scale and the items. Since each of the five sets of samples targeted one of the five rating categories, raters were given the opportunity to try differentiating performances against all five IC categories. Table 45 below illustrates how the second step of rater training was designed.

Table 45 Content in the second step of rater training

The sub-scale used	**Disaffiliation control**	**Affiliation promotion**	**Morality**	**Reasoning**	**Role enactment**
The items chosen	Eight sample performances on item 9_ complaining to your boss	Eight sample performances on item 7_ criticizing friend's son	Eight sample performances on item 4_ breaking bad news to a neighbour	Eight sample performances on item 8_ disagreeing with your classmate	Eight sample performances on item 5_ criticizing your team member

Step three: **Mock rating** using pilot test-taker performances. The last step in rater training involved the actual rating of test-taker performances using all five IC categories in the rating scale. I pre-selected 18 sample performances from pilot test-takers covering all nine test items. The samples were selected based on the representativeness of the differing IC abilities in pilot test-takers, relying on the Rasch fair scores from pilot rating results. Raters were asked to listen to the 18 sample performances and give five scores corresponding to the five IC

categories. They were encouraged to consult the CA-informed exemplars in Step one and the samples used in Step two to assist them with deciding which band/score on each of the five categories was appropriate. After the raters finished rating, I conducted a moderation meeting to examine their ratings compared to mine. There was high consistency between the scores the two raters assigned, demonstrated by the high Spearman correlation between their scores ($r=0.79$, $p<0.05$). High rater agreement was also achieved between the two raters' and my rating. Out of 90 ratings (18 samples * 5 categories), the number of exact agreements for all three raters, which included the researcher, was 23 (25.6 %). The number of cases where there was a one-point difference among the three raters was 46 (51.1 %) while the number of two- and more-than-two-point differences was 21 (23.3 %). In the post-rating mediation session, I focused on the ratings that attracted two-point-or-above difference and discussed with the raters to understand why their ratings differed from the ones of each other's and mine. Consensus was achieved for all the cases discussed.

Once the three steps of rater training were completed, raters confirmed they felt confident and comfortable with using the IC rating scale to rate test-taker performance. They also acknowledged the facilitative effect of CA-informed exemplars and other training materials to assist them to develop better insight into the scale and achieve consistency in rating.

6.1.3.3 Rater rating

Having completed rater training and reached the desirable consistency in pilot rating, both raters started rating the test performances of the 105 test-takers. I adopted a fully-crossed rating scheme to examine rater reliability. This means every rater rated every test-taker's performance on each of the nine items on each of the five IC rating categories, totalling 4725 ratings for each rater. The rating scheme was to ask raters to rate test-takers' performance by items, which means that raters first rated all 105 test-takers' performances on item 1, and then all 105 performances on item 2, until all 105 performances on item 9. The purpose of this design was that raters could focus on one item each time instead of having to constantly shift between different items if they were asked to rate performances by test-takers. This design could also potentially lower the halo effect among the nine items from the same test-taker.

6.1.4 Data analysis

Once raters finished ratings, I collected raters' rating results and formatted them for FACETS Rasch analyses. Questionnaire data were coded with 1 given to

strongly disagree, 2 to *disagree*, 3 to *slightly disagree*, 4 to *slightly agree*, 5 to *agree* and 6 to *strongly agree*. Negatively stated items were reverse-coded. If answered *unable to assess*, that item was treated as missing data.

6.2 Results and initial discussion

Section 6.2 presents the findings of study three, focusing on the IC test results, the questionnaire results and the relationships between test scores and questionnaire scores. Rasch analysis of the rating scale functioning was presented at the 2021 American Association for Applied Linguistics (AAAL) conference (Dai, 2021).

6.2.1 Rasch analyses of IC test scores

6.2.1.1 The Wright map

The reporting of the results for study three starts with the Rasch analyses of the IC test scores. Many Facet Rasch generates a range of useful reports on the functioning of items, raters, test-takers, and steps in the rating scale. First, let us inspect the Wright map for this dataset.

Wright maps in Rasch are a useful tool that expresses test-taker ability, relative rater severity, item difficulty and criterion/rating category[14] difficulty on the same logit scale. The Wright map in this study provides a general picture of whether the raters identified test-takers' differing IC. Figure 21 is the Wright map showing all 105 test-takers' performances on the nine items in the IC test, rated by two raters.

The candidate column provides information regarding test-takers' candidate ID, which goes from 1 to 105 for all 105 test-takers. The item column offers information about the specifications of the items by listing their Power variable, language use domain and task delivery method. The criteria column shows the five rating categories in the rating scale. Finally, the scale column displays raters' use of the five steps in the rating scale. This Wright map was derived from the rating scale Rasch model, which treats the five sub-scales from the five rating categories as one overall scale. The average rater severity, item difficulty and criterion difficulty were set as zero while test-takers were allowed to float.

From Figure 21 we can see the raters spread test-taker abilities across five logits, which shows that raters were capable of identifying differences in test-taker performances using the rating scale. The distribution of test-takers in

14 In FACETS Rasch analysis, a rating category in an analytic rating scale is referred to as a criterion.

general conformed with a normal distribution, with test-takers stronger in IC mapped towards the positive end of the logit scale and test-takers weaker in IC towards the negative end of the logit scale. The rater column shows that the two raters were similar in their severity as their positions were equal in the Wright map. A detailed discussion on inter-rater reliability will be offered later but here it is sufficient to say the two raters were similar in their assessment of test-taker ability using the IC scale for this test. The item column reveals that most items were within one logit of difficulty except for item 5, a work-domain second pair-part voice message item where the test-taker supposedly had more power than their interlocutor. Item 5 was noticeably more difficult than others, evidenced by its higher position than other items on the logit scale. In terms of the five rating categories, *social role management* was the most difficult category as it was the highest in the criteria column. This means it was the hardest rating category for test-takers to score high on. *Affiliation promotion* and *reasoning* were similar in difficulty, while *disaffiliation control* and *morality* were easier to perform well on. Finally, the scale column demonstrates that the raters effectively used the five steps in the rating scale as all five steps were represented on the logit scale, with the highest and lowest steps in brackets.

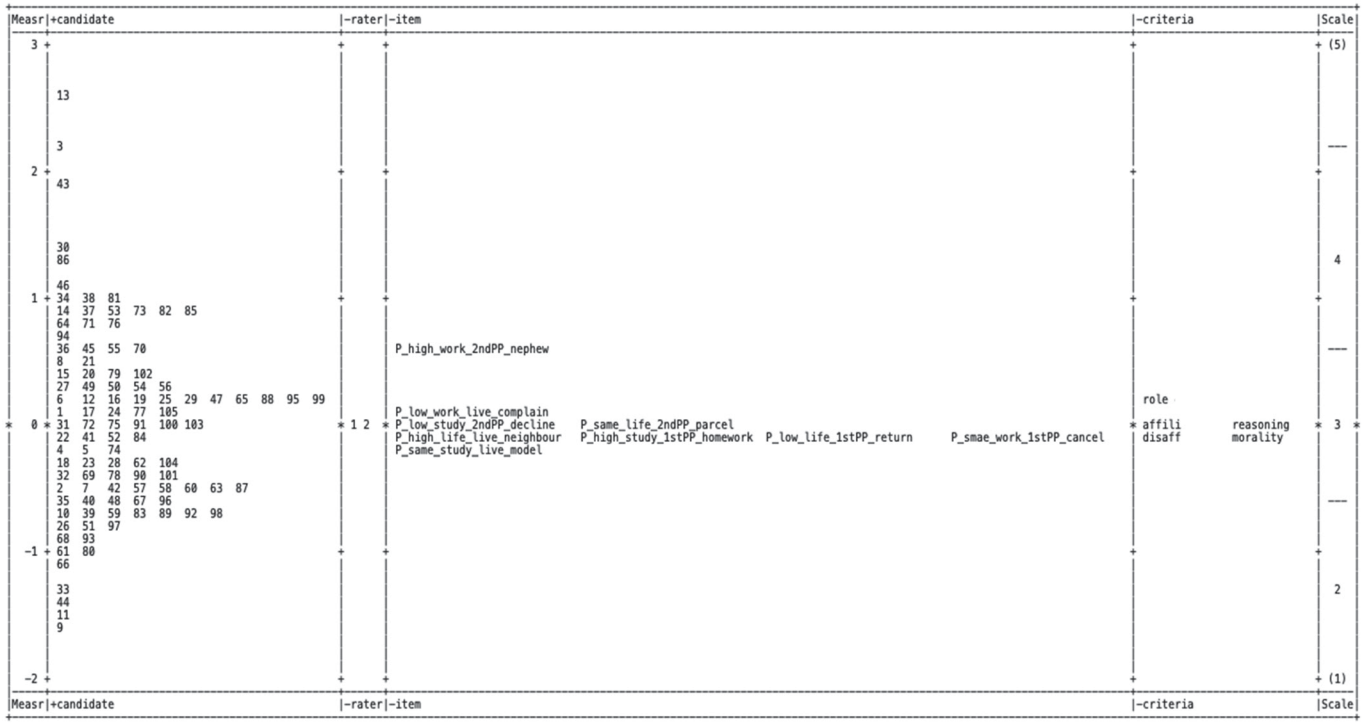


Figure 21 Wright map from the main testing dataset

In summary, the Wright map offers a snapshot of raters' ratings using the IC rating scale. Information from the Wright map attests to the statement that raters used the rating scale successfully to differentiate differing test-taker performance, taking into account the differences between item and criteria difficulties.

6.2.1.2 The candidate measurement report

Many-facet Rasch generates a candidate measurement report that illustrates if raters can use the scale to separate test-takers to a range of ability levels as expected by the test developer. The Wright map discussed in Section 6.2.1.1 already shows that there was a wide spread of test-taker ability levels in the candidate column. More information regarding the separation of test-takers can be found in the Rasch candidate measurement report. Due to the length of the report, it is not reproduced here but can be found in its entirety in Dai (2022). Here Table 46 presents the summary statistics of the candidate measurement report.

Table 46 Summary statistics of the candidate measurement report

Model, Populn: RMSE .12 Adj (True) S.D. .73 Separation 6.02 Strata 8.35 Reliability .97
Model, Sample: RMSE .12 Adj (True) S.D. .73 Separation 6.05 Strata 8.39 Reliability .97
Model, Fixed (all same) chi-square: 3278.3 d.f.: 104 significance (probability): .00
Model, Random (normal) chi-square: 100.6 d.f.: 103 significance (probability): .55

The candidate measurement report in Dai (2022) shows that the observed average score range for the test-taker population was 1.8–4.61 on the original 1–5 scale (mean=3.01, SD=0.56). Rasch also produces fair scores for test-takers, which are corrected for differences in rater severity. The fair average score range for this study was 1.79–4.62 (mean=3.01, SD=0.57). Looking at the complete candidate measurement report in Dai (2022) we can see that there was little difference between observed averages and fair averages. This was due to the fully crossed rating design adopted in this study. Since every test-taker's every performance on every test item was rated by both raters, the average rating results already took into account the differences in the two raters' ratings.

The measure column in the candidate measurement report details the estimate of test-taker ability in logit and the model SE column shows how accurate the estimate is. The range of ability measures for the test-taker population in this study was from -1.64 to 2.56, corresponding to the more-than-4-logit spread in the Wright map. The model SE was small (mean=0.12, SD=0.01), which was due to the copious amount of information available regarding test-takers' ability,

since every test-taker's performance on each of the nine items was rated by two raters. The results indicate the Rasch programme confirmed that test-takers' ability was measured with precision.

Candidate fit statistics in the report provide information about the consistency of rater rating as it looks at how predictable a candidate's score is in relation to the whole test-taker population (McNamara et al., 2019). Here the infit statistic was investigated as it was not influenced by outliers. One method McNamara et al. (2019) detailed for calculating the appropriate range for infit statistics is to use mean plus and minus twice the S.D of the mean square statistic (MnSq reported in Dai, 2022). Adopting this method, I calculated the range for this test-taker population, which was 0.4–1.6 (mean=1.0, SD=0.3). Test-takers with an infit statistic of above-1.6 meant their scores did not conform with the expectation of the Rasch model. This is not a concern for this dataset as only three (test-taker ID 37, 6 and 100) out of 105 test-takers were misfitting.

The summary statistics in the candidate report in Table 46 provide useful information in relation to whether the test separated test-takers to a sufficient number of ability levels. A 6.05 separation ratio indicated the test and rating scale separated test-takers to a wide range of abilities with precision, while an 8.39 strata showed that the test and rating scale differentiated 8.39 groups of abilities in the test-taker sample. These findings support the evaluation inference in the validity argument as they demonstrate the rating scale functions well in distinguishing test-takers' of differing abilities.

6.2.1.3 The rater measurement report

The rater measurement report from Rasch generates useful information about rater behaviour and rater reliability. The rater measures are the logit position of raters in the Wright map, indicating rater severity. Table 47 presents the rater measurement report for this study and shows that the two raters for this test had very similar measures, which were 0.02 and -0.02 respectively, with a 0.02 SE. The difference between measures was only 0.04 logit. This shows that both raters rated in a very similar manner, demonstrating comparable severity and leniency. The low SE supports the argument that the estimate of the rater measure was precise.

Similar to the fit statistics in the candidate measurement report, the fit statistics for raters are indicators of how predictable raters' ratings are considering the whole dataset, which is a measure of rater consistency. Adopting the same range criterion (mean+/- 2 SD), the acceptable infit range for raters was 0.94–1.06. Both raters had infit values (1.02 and 0.98) within this range, demonstrating that

the raters were capable of using the rating scale to rate in a consistent manner, without unexpected harshness or leniency in their rating. Both raters were also not overfitting (infit values under 0.94), showing little halo effect or central tendency effect (McNamara et al., 2019).

In terms of the separation and strata values for raters, both values were very small for this sample (1.65 and 2.54 respectively). This suggests little difference in the manners the two raters rated in the test, pointing to high inter-rater reliability. A moderate reliability index (0.73) shows that Rasch was not confident in distinguishing the differences in the two raters' ratings, which is another piece of evidence supporting strong inter-rater consistency.

Lastly, the last line in Table 47 shows the number of exact agreements between the two raters in their awarding of scores. Out of 4725 opportunities for agreement, the two raters agreed in their rating 2208 times, which was 46.7 % out of the total amount. This is relatively high considering there were five steps in the rating scale for the raters to choose from in each rating.

Table 47 The rater measurement report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N rater
14122	4725	2.99	3.00	.02	.02	1.02	1.0	1.02	1.1	.97	.55	.56	46.7	31.9	1 1
14287	4725	3.02	3.03	-.02	.02	.98	-1.0	.98	-.9	1.03	.57	.56	46.7	31.9	2 2
14204.5	4725.0	3.01	3.01	.00	.02	1.00	.0	1.00	.1		.56				Mean (Count: 2)
82.5	.0	.02	.02	.02	.00	.02	1.1	.02	1.1		.01				S.D. (Population)
116.7	.0	.02	.03	.03	.00	.03	1.5	.03	1.5		.02				S.D. (Sample)

Model, Populn: RMSE .02 Adj (True) S.D. .02 Separation .93 Strata 1.58 Reliability (not inter-rater) .46
Model, Sample: RMSE .02 Adj (True) S.D. .03 Separation 1.65 Strata 2.54 Reliability (not inter-rater) .73
Model, Fixed (all same) chi-square: 3.7 d.f.: 1 significance (probability): .05
Inter-Rater agreement opportunities: 4725 Exact agreements: 2208 = 46.7% Expected: 1505.1 = 31.9%

6.2.1.4 The criterion measurement report

The criterion measurement report presents the difficulty measures and fit statistics for the five IC rating categories. As already indicated in the Wright map, the category social role management was more difficult than the other four categories. The criterion measurement report provides more detailed information regarding the difference in difficulty measures for each of the five categories. From the measure column in Table 48, it is observable that *social role management* was 0.23 logit more difficult than the second most difficult category *reasoning*. *Reasoning* and *affiliation promotion* were similar in difficulty with 0.01 logit for the former and -0.02 logit for the latter. It was relatively easier for test-takers to achieve higher scores on *morality* and *disaffiliation control* as both had lower logit values (-0.10 and -0.12). This conforms with the theoretical expectation discussed in Section 5.2.6 as *social role management*, a higher-order

IC indicator hypothesized in this book that requires overarching abilities in managing social interaction, was speculated to be more challenging to perform well on compared to other IC rating categories (see how *social role management* was positioned on top of the other four categories in Figure 20). It is plausibly easier for test-takers to score higher on *morality* and *disaffiliation control* because the majority of test-takers, who were everyday members of society, would try to present themselves as morally positive in testing settings. It is also commonsensical for them to want to mitigate disaffiliation in performance assessment settings where the test tasks required them to perform disaffiliative actions. The differently positioned, but also theory-confirming five categories, therefore, offered holistic and comprehensive coverage of IC as a test construct.

Using the same method for infit range calculation, the range for the rating categories was 0.88 to 1.12. All five criteria/rating categories were neither overfitting nor misfitting. This provided preliminary evidence that the five IC categories measured a unidimensional IC construct. The unidimensionality of the test was further interrogated using other Rasch statistical tests, which will be reported later.

Similar to the analyses for previous measurement reports, the separation and strata values (5.4 and 7.53 respectively) for the rating categories evince that the five categories were able to be separated to 7.53 distinct difficulty levels. This demonstrates that the categories were clearly dissimilar in their difficulty and measured the IC construct at different levels. High separation reliability (0.97) was also achieved, bolstering confidence in the difference in the difficulty measures of the five categories. This outcome is desirable as the rating categories should have different difficulty levels so as to contribute unique information, but at the same time measure the same IC test construct.

Table 48 The criterion measurement report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N criteria
5333	1890	2.82	2.82	.24	.03	1.00	-.1	.99	-.2	1.01	.56	.56	5 role
5673	1890	3.00	3.01	.01	.03	1.05	1.6	1.05	1.5	.94	.55	.55	4 reasoning
5716	1890	3.02	3.03	-.02	.03	1.01	.2	1.02	.6	.99	.56	.55	2 affili
5828	1890	3.08	3.10	-.10	.03	.90	-3.2	.92	-2.8	1.09	.50	.55	3 morality
5859	1890	3.10	3.12	-.12	.03	1.04	1.2	1.03	1.1	.97	.58	.55	1 disaff
5681.8	1890.0	3.01	3.01	.00	.03	1.00	.0	1.00	.0		.55		Mean (Count: 5)
187.4	.0	.10	.11	.13	.00	.05	1.7	.05	1.6		.03		S.D. (Population)
209.6	.0	.11	.12	.14	.00	.06	2.0	.05	1.8		.03		S.D. (Sample)

Model, Populn: RMSE .03 Adj (True) S.D. .13 Separation 4.81 Strata 6.74 Reliability .96
Model, Sample: RMSE .03 Adj (True) S.D. .14 Separation 5.40 Strata 7.53 Reliability .97
Model, Fixed (all same) chi-square: 120.2 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 3.9 d.f.: 3 significance (probability): .27

6.2.1.5 The item measurement report

The item measurement report offers insight into the functioning of the nine items in the test. The measure column in Table 49 expresses item difficulty in logits and the nine items were within one logit of difficulty, ranging from −0.22 to 0.57. The small SE for item measures (0.4) indicates that item difficulty was measured with precision, due to the relatively large sample size and the fully-crossed rating design in this study. The infit statistic range for items was 0.76–1.24 (mean=1, SD=0.12). All items were within this range, neither overfitting nor misfitting. The PtMea Correlation column provides a classic test theory measure of the correlation of each item's raw score with the total raw score of the rest of the items. All nine items had moderate PtMea correlations, indicating again that they were neither misfitting nor overfitting. This is evidence that all the items were measuring the same construct since if some of the items were measuring a different construct, their PtMea correlation values would be negative (Wright, 1992). This was not the case for the nine items in this study. The fact that all items were not misfitting and the PtMea correlation values were positive offers additional pieces of evidence demonstrating that this test was measuring a unidimensional construct.

Finally, different from the candidate/test reliability index in the candidate measurement report, the item reliability index is an indicator of how generalizable the ordering of items is if the items were given to different cohorts of test-takers (McNamara et al., 2019). A high item reliability index, as found in this study (0.98), provides backing for the generalization inference in that different groups of test-takers are likely to find the order of items, in terms of item difficulty, in the same order as in this study.

Table 49 The item measurement report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	item
2695	1050	2.57	2.54	.57	.04	1.23	5.3	1.21	4.7	.74	.58	.55	5	P_high_work_2ndPP_nephew
3066	1050	2.92	2.92	.11	.04	1.06	1.4	1.05	1.1	.93	.54	.55	9	P_low_work_live_complain
3154	1050	3.00	3.01	.00	.04	.83	−4.4	.83	−4.3	1.21	.58	.54	4	P_same_life_2ndPP_parcel
3166	1050	3.02	3.02	−.01	.04	1.00	.0	1.00	.0	1.00	.55	.54	6	P_low_study_2ndPP_decline
3225	1050	3.07	3.08	−.08	.04	.91	−2.3	.91	−2.3	1.12	.58	.54	2	P_low_life_1stPP_return
3243	1050	3.09	3.10	−.11	.04	.88	−3.1	.90	−2.5	1.14	.55	.54	1	P_high_study_1stPP_homework
3252	1050	3.10	3.11	−.12	.04	1.09	2.0	1.09	2.2	.87	.42	.54	3	P_smae_work_1stPP_cancel
3274	1050	3.12	3.13	−.15	.04	.97	−.7	.97	−.7	1.06	.57	.54	7	P_high_life_live_neighbour
3334	1050	3.18	3.20	−.22	.04	1.04	1.0	1.06	1.4	.94	.50	.54	8	P_same_study_live_model
3156.6	1050.0	3.01	3.01	.00	.04	1.00	−.1	1.00	.0		.54		Mean (Count: 9)	
178.8	.0	.17	.18	.22	.00	.11	2.8	.11	2.7		.05		S.D. (Population)	
189.7	.0	.18	.19	.24	.00	.12	3.0	.11	2.8		.05		S.D. (Sample)	

Model, Populn: RMSE .04 Adj (True) S.D. .22 Separation 6.23 Strata 8.65 Reliability .97
Model, Sample: RMSE .04 Adj (True) S.D. .23 Separation 6.62 Strata 9.16 Reliability .98
Model, Fixed (all same) chi-square: 350.0 d.f.: 8 significance (probability): .00
Model, Random (normal) chi-square: 7.8 d.f.: 7 significance (probability): .35

6.2.1.6 The rating scale category functioning

The rating scale statistics produced by Rasch generate information regarding how the rating scale was utilized by raters. Following the guidelines proposed by Linacre (2002), Table 50 shows that there were more than 10 observations in each scale step, ensuring stable estimates. The average measures were advancing at similar values, which were 0.31, 0.34, 0.38 and 0.54 logits between the five steps. It is understandable that band 5, exemplary, had a relatively higher value distance from band 4 because band 5 was supposed to measure performance that was truly outstanding. The average measures were also advancing monotonically and not disordered. The distribution of data points across the five steps approximated a normal distribution. Expected measures were also similar to average measures, with differences smaller than 0.04 logit. The mean-square values ranged from 0.9 to 1.1, which were below 2 and conformed with the expectations in Linacre (2002). The Rasch-Andrich thresholds measures were advancing at 1.2, 1.09 and 1.22, which were monotonical. It should be noted that since the rating scale in this study was not conceptualized as a set of dichotomous items, step difficulties did not need to advance by at least 1.4 logits in this case (Linacre, personal communication).

Having examined the scale functioning against the guidelines in Linacre (2002), it is safe to conclude that the IC scale functioned as expected and raters adequately distinguished the different steps in the scale.

Table 50 Rating scale step statistics

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	719	719	8%	8%	-.70	-.74	1.1			(-3.00)		low	low	100%
2	2345	2345	25%	32%	-.39	-.39	1.0	-1.75	.04	-1.29	-2.22	-1.75	-1.98	46%
3	3313	3313	35%	67%	-.05	-.02	1.0	-.55	.02	.00	-.61	-.55	-.58	42%
4	2304	2304	24%	92%	.43	.40	.9	.54	.02	1.30	.61	.54	.57	46%
5	769	769	8%	100%	.97	.96	1.0	1.76	.04	(3.03)	2.23	1.76	1.98	100%
										(Mean)		(Modal)	(Median)	

Figure 22 below presents the rating scale probability curves which show the probability of a test-taker at a given logit ability score (x axis) getting a particular band score (the curves). As Figure 22 illustrates, each step in the rating scale had a separate curve and a distinct peak. This indicates raters differentiated different steps in the scale and used the scale as expected by the scale developer.

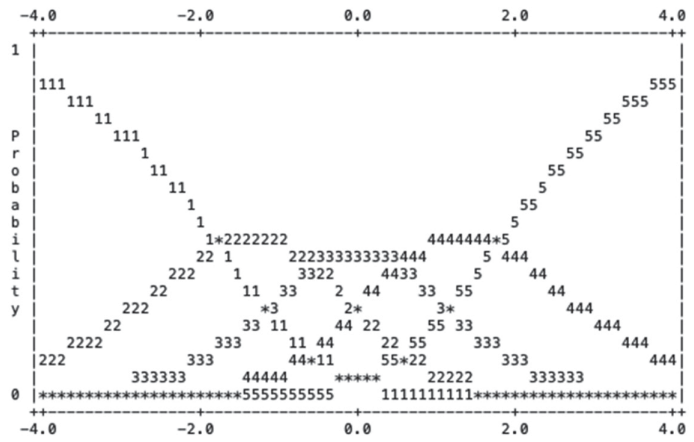


Figure 22 Rating scale probability curves

6.2.1.7 The dimensionality of the data structure

Finally, a Rasch residuals principal components analysis (PCA) was conducted on the rating results to examine if there was more than one variance component in the data structure. Wright (1996) argued that if the data is unidimensional, the components in the residuals should be at the noise levels. In this PCA, residuals were standardized by their model SD as suggested in Linacre (1998). There were five items in this PCA, corresponding to the five rating categories of standardized residuals. Each of the five items was modelled to contribute one unit of unexplained variance.

The results in Table 51 show that the total raw variance was 56.02 units (eigenvalue), 51.02 units of which were explained by measures. Within the 51.02 units, 18.10 units were explained by persons and 32.91 units by items, which were the rating categories in the IC test. This shows that the test-takers and the rating categories explained 91.1 % of the variance in the dataset.

Table 51 Standardized residual variance in PCA

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed	Expected
Total raw variance in observations =	56.0163	100.0%	100.0%
Raw variance explained by measures =	51.0163	91.1%	88.6%
Raw variance explained by persons =	18.1052	32.3%	31.4%
Raw Variance explained by items =	32.9111	58.8%	57.1%
Raw unexplained variance (total) =	5.0000	8.9% 100.0%	11.4%
Unexplned variance in 1st contrast =	1.5257	2.7% 30.5%	
Unexplned variance in 2nd contrast =	1.3442	2.4% 26.9%	
Unexplned variance in 3rd contrast =	1.1038	2.0% 22.1%	
Unexplned variance in 4th contrast =	1.0237	1.8% 20.5%	
Unexplned variance in 5th contrast =	.0022	.0% .0%	

The focus here is the variance unexplained by the model, which was 5 units and accounted for 8.9 % of the total variance. PCA decomposed the inter-item correlations and discovered that the first contrast (the first component in the PCA) explained 1.5 units of the variance. As each of the five items/rating categories was modelled to explain one unit of the variance, 1.5 units were slightly higher than the ideal value of one unit. However, this is not uncommon for the first contrast since the first contrast always exceeds the eigenvalue of 1.4 units. Therefore, there is nothing suggestive of additional dimensions based on the first contrast.

If we look at the relationships between the person measures on the first three clusters of items in Table 52, their correlations were close to 0.9 or above, which is very high. It is usually when the inter-person correlations are less than 0.7 that different dimensions become a concern (Linacre, personal communication). Therefore, the high correlations here did not suggest the existence of multidimensionality.

Table 52 PCA relationships between the person measures

Approximate relationships between the PERSON measures

PCA Contrast	ITEM Clusters	Pearson Correlation	Disattenuated Correlation	Pearson+Extr Correlation	Disattenuated+Extr Correlation
1	1 – 3	0.8780	0.9637		
1	1 – 2	0.8928	0.9868		
1	2 – 3	0.9407	1.0000		
2	1 – 3	0.8973	1.0000		
2	1 – 2	0.9415	1.0000		
2	2 – 3	0.9482	1.0000		
3	1 – 3	0.9415	1.0000		
3	1 – 2	0.9482	1.0000		
3	2 – 3	0.8973	1.0000		
4	1 – 3	0.9519	1.0000		
4	1 – 2	0.9212	1.0000		
4	2 – 3	0.9202	1.0000		
5	1 – 3	0.8973	1.0000		
5	1 – 2	0.9482	1.0000		
5	2 – 3	0.9415	1.0000		

Considering the results of the Rasch PCA, there was no evidence of multidimensionality in the five items/rating categories in the IC test dataset. This suggests that there is only one variance component, test-taker IC, that explains the structure of the main testing dataset.

6.2.2 Correlation between IC and LC

Having analysed the 105 test-takers' test scores on the IC test using Rasch in Section 6.2.1, Section 6.2.2 examines the relationship between test-takers' IC and linguistic competence (LC). The relationship between IC and LC has long

engaged the attention of IC assessment researchers (Ikeda, 2017; Ockey et al., 2015; Roever & Ikeda, 2021; Youn, 2013) and has been explored in depth in Sections 2.3 and 2.4 in the literature review. For the 105 test-takers in study three, their HSK scores were used as a measure of their LC.

HSK, the Chinese Standard Exam, is a standardized and validated large-scale proficiency test for L2 speakers of Chinese (Peng et al., 2020). The test claims to assess L2-Chinese speakers' ability in 'using the Chinese language in their daily, academic and professional lives' (Chinesetest, n.d.). HSK has six levels from HSK1 to HSK6 of increasing proficiency. HSK test developers have linked HSK levels to the CEFR levels, with HSK1 linked to CEFR A1, up to HSK6 to CEFR C2. Although HSK only has listening, reading, and writing sections without a speaking section, test developers of HSK argue that it can assess test-takers' communicative competence in real-world language use (Chinesetest, n.d.). Both the linking claim and the real-world extrapolation claim have been subjected to queries and more research is needed to assess the validity evidence for these claims (Peng et al., 2020). Nevertheless, HSK remains a validated large-scale Chinese proficiency test with high uptake among end-users and can serve as a valid criterion measure of test-takers' proficiency/LC.

Considering the concerns mentioned earlier in regard to the construct representation in HSK, test-takers' HSK scores in this study were only used as an indicator of their general LC, instead of a precise measure of their LC or speaking ability (a discussion on the difference between speaking/LC and talking/IC can be found in Sections 2.3 and 2.4). In other words, test-takers' HSK scores were viewed as indicative of the linguistic devices they possessed in mobilizing language to facilitate communication, without any claims being made in this book about their ability to use language in the speaking mode successfully or appropriately.

Another limitation of using HSK as a measure of LC is that HSK scores are scaled to band categories. Continuous measures of LC would be more advantageous in terms of conducting correlational analyses of the relationship between LC and IC, although it would entail an investment of resources that go beyond the constraints of this book. The use of HSK scores to measure test-takers' LC, therefore, is a less-than-ideal but nonetheless necessary compromise considering limited resources I had when conducting this study.

104 out of all 105 test-takers had taken HSK prior to participating in the main testing session in study three and Table 53 details the number of test-takers in each HSK group, the mean of their IC test Rasch fair scores, and the SD of the fair scores for each group. Figure 23 is a visual representation of the distribution of test-takers' HSK levels (x axis) against their fair scores (y axis). In Figure 23 HSK 7 is a shorthand for the L1-Chinese group (N=15).

Table 53 Test-takers' HSK levels and IC fair scores

	<HSK3	HSK3	HSK4	HSK5	HSK6	L1-speaker
N	7	12	11	29	30	15
Mean	2.66	2.79	2.51	2.96	3.21	3.34
SD	0.30	0.37	0.38	0.46	0.51	0.80
Min	2.23	2.16	1.89	2.37	1.79	1.96
Max	3.03	3.32	3.19	3.98	4.06	4.62

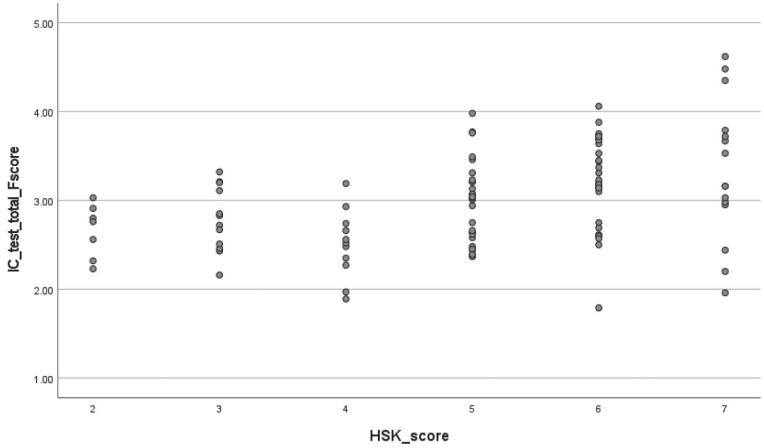


Figure 23 Test-takers' HSK levels and IC fair scores plotted

Descriptive statistics and visual representation show that as test-takers' proficiency increased by HSK levels, their IC average scores in general also increased, with growingly larger SDs. It is worth noting that test-takers from the HSK4 group had a lower mean than the means from the below HSK3 and HSK3 groups. This could be explained by the U-shaped learning in pragmatics (Polat, 2011; Taguchi, 2012; Timpe-Laughlin, 2017) or by the sampling of test-takers. A more revealing observation is that as test-takers' proficiency improved, there was more variation and dispersion in their IC test performance, evidenced by the increase in SD. Most noticeably, the L1-Chinese group had the widest range of performance, with an SD of more than 2.5 times larger than the SD of the below HSK3 group. This supports the argument that higher proficiency/LC is unable to guarantee stronger or more homogenous IC performance as L1 speakers, by default, should have the strongest LC. One explanation is that IC is dependent on

LC but cannot be solely predicted by LC. Put in another way, in order to achieve high IC, speakers need to have a reasonable level of LC so as to have the linguistic resources to implement social actions. This is evidenced by the fact that no test-takers below HSK4 were able to achieve a mean IC fair score higher than 3.5. This point is also corroborated by the fact that only HSK6 and L1 test-takers achieved above 4 fair scores. A high degree of LC is needed to tailor interaction with the desired IC features.

Although strong LC is a necessary condition for strong IC, strong LC cannot guarantee strong IC. In fact, the lowest fair scores in the HSK6 and L1 groups were lower than the ones in the below HSK3 and HSK3 groups. This shows that despite the stronger LC possessed by test-takers who were at the HSK6 proficiency level or who were L1-Chinese speakers, they could still err on the IC side, and sometimes could commit more serious IC blunders than test-takers with less LC. A possible explanation for this phenomenon is that when L2 speakers first acquire an L2, they learn the interactional and pragmatic norms from textbooks and teachers, modelling their interaction and speech on exemplars that conform with general social expectations. This explains the narrower ranges of their fair scores. As their proficiency increases, they start to have more linguistic resources to embody their attitudes, beliefs, and values towards interaction in the L2 context, qualities that typify Hymes's model of communicative competence (see Section 2.4.3) and that constitute the construct of IC in the IC test in this book (see Section 5.2.6). When such attitudes, beliefs and values concur with the normative expectations in the target language community, their interaction was viewed favourably, reflected in higher IC scores. When their attitudes, beliefs and values were not aligned with the social norms in the target community, they were perceived to be having less IC, though their LC could be much stronger than when they were less proficient in their use of the language. This was most starkly reflected in the L1-Chinese group, where the widest range of IC performances was found. L1-Chinese speakers do not form a homogenous group where every member is strong in IC. Conventional wisdom informs us that L1 speakers have a wide range of IC as some L1 speakers are stronger at interpersonal communications while others are not. The fact that there was a wide dispersion of IC scores within the L1 speaker group is strong backing for the separation of IC and LC at higher ability levels.

Another factor that can contribute to the difference in SD is that L2 speakers were more cognizant of the assessment nature when they were performing the role plays in the test, due to the fact that they were using a language they had learnt additional to their L1. L2 speakers might have a propensity to control their performance to conform with the interactional norms they believed that

are prevailing in the L2 culture. L1 speakers, on the other hand, might have a stronger sense of agency and ownership of the language and hence displayed less scrupulousness in their use of the language. To them, they were doing *being themselves* in the test tasks, compared to doing *performing a perceived model speaker* of the language which might be the primary goal for L2 speakers in assessment contexts.

One counterargument to the abovementioned analysis is that the number of participants in each HSK group was not the same as there were fewer HSK 3 and HSK4 test-takers. This could have artificially dampened the variability in those two groups. To address this concern, Table 54 presents a re-group of test-takers into four groups, consisting of a group for test-takers at or below HSK4, a group at HSK5, a group at HSK6 and an L1-speaker group. This grouping ensured similar numbers of test-takers in the three L2-Chinese speaker groups. The trends in terms of variability (SD), means, minimum scores and maximum scores in the groups, as observed in Table 53, still held in Table 54. The variability/SD in IC increased as test-takers' LC/HSK went up, supporting the arguments established earlier about the relationship between IC and LC.

Table 54 Test-takers' HSK levels and IC fair scores (regrouped)

	Up to HSK 4	**HSK5**	**HSK6**	**L1-speaker**
N	30	29	30	15
Mean	2.65	2.96	3.21	3.34
SD	0.37	0.46	0.51	0.80
Min	1.89	2.37	1.79	1.96
Max	3.32	3.98	4.06	4.62

In terms of the correlation between test fair scores/IC and HSK scores/LC, Spearman's rho showed that the two variables correlated at 0.416 ($N=104$, $p<0.05$). The correlation between the IC test and proficiency in this study was smaller than the ones in previous IC assessment research (0.8 in Ikeda, 2017; 0.9 in Youn, 2013). This is understandable when the proficiency measure in this book, HSK, was not directly measuring speaking. Another possible explanation is that the IC construct measured in this book covers a broader range of sequential, moral, logical, emotional, and categorial IC indicators (see Section 5.2.6), which departs more noticeably from the LC indicators captured in HSK.

6.2.3 Rasch analyses of questionnaires

Having examined the functioning of the IC test and its connection with LC in Section 6.2.1 and Section 6.2.2, Section 6.2.3 looks at the accompanying self-assessment and peer-assessment IC questionnaires. As mentioned in Section 6.1.2.3 in the methodology section, test-takers were required to finish a self-assessment questionnaire on their self-perception of their IC. They were also encouraged to find an L1-Chinese peer whom they interacted with frequently in Chinese to complete a peer-assessment questionnaire so their peers could assess test-takers' IC from an L1 peer perspective based on test-takers' IC *in the wild*. Both self and peer-assessment questionnaires contain the same 35 items targeting the five rating categories in the rating scale. The questionnaires generated insight into how IC was perceived from the L2-Chinese test-taker and L1-Chinese interactant perspectives. Such insight also provided an understanding of how much test-takers' test performance on the IC test mirrored their and their peers' perceptions of their IC in non-testing settings. Here Rasch analysis was utilized to analyse data from both questionnaires.

Rasch is a suitable technique for analysing questionnaire data when the questionnaire items are designed to measure the same underlying construct, which in this case, is test-takers' perception of their IC for the self-assessment questionnaire and test-taker peers' perception of test-takers' IC for the peer-assessment questionnaire. Compared to using solely descriptive statistics to analyse questionnaire data, Rasch offers a clear representation of how the items are measuring different levels of the construct and how the questionnaire items differ in logit values based on the difficulty in endorsing or dis-endorsing the propositions in the questionnaire items (McNamara et al., 2019). Here the rating scale model was used since all 35 items had the same step structures (six response options from *strongly disagree* to *strongly agree*). Both the self and peer-assessment questionnaires can be found in Appendix V and VI. In the following sections, I will first present the results of the five sub-sections of the questionnaires, which correspond to the five rating categories, before offering a discussion on all the items in the questionnaires.

6.2.3.1 The disaffiliation control sub-section

Two separate Rasch analyses were conducted on the eight *disaffiliation control* items in both the self and peer-assessment questionnaire data. Item 35 turned out to be misfitting in both cases, so it was excluded from this sub-section. Rasch analyses were then rerun on the remaining items and showed that all seven items were within the infit range for items for both questionnaires. Using the mean +/

- 2 SD criterion, the infit range for items in the self questionnaire was 0.61–1.37 while the range for items in the peer questionnaire was 0.68–1.24.

Rasch test reliability for the self-assessment questionnaire was 0.77 while for the peer-assessment questionnaire it was 0.80. High test reliability indexes supported the argument that both sub-sections were measuring unidimensional constructs. Table 55 presents the items in the questionnaire that targeted the ability in *disaffiliation control*. The propositions in the items, the fair averages for both the self and peer questionnaires and their SE were provided. The infit values for each item in both questionnaires were also listed to verify that they were within their respective infit ranges. RC indicates the item was reverse-coded.

Table 55 Disaffiliation control questionnaire item analysis

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
2	Can choose language to avoid offending others	4.54	0.11	0.81	5.35	0.20	0.79
8	Can avoid sounding direct or forceful (RC)	3.35	0.10	1.15	4.30	0.14	0.98
15	Can talk in a calm and collected manner when one can potentially offend others	4.56	0.11	0.76	5.20	0.19	1.13
17	Can control one's negative emotions when conversing	4.44	0.11	1.17	5.41	0.21	0.81
21	Can use indirect language to manage difficult interpersonal situations	4.29	0.11	0.79	5.15	0.18	0.81
22	Can avoid rejecting/ disagreeing in an impolite manner (RC)	3.75	0.11	1.26	4.70	0.15	1.03
31	Never use language that is overly direct or forceful	4.05	0.10	1.00	5.10	0.17	1.15
35	Can avoid sounding threatening in Chinese (RC)	Misfitting	/	/	Misfitting	/	/

The comparison between the fair average scores for each item in this sub-section between self-assessment and peer-assessment highlighted the similarities and differences in test-takers' and their peers' perceptions of IC at different item levels. It is clear that test-taker peers viewed test-taker performance *in the*

wild more favourably than test-takers did, with on average one point higher in average fair scores. Both groups noticed that test-takers were relatively stronger at adopting linguistic devices to mitigate disaffiliation (item 2), maintain a placid demeanour when working with disaffiliative situations (item 15) and control their emotions to rein in disaffiliative impacts on social relations (item 17). Two negatively phrased items, item 8 and item 22, elicited the opinions from both groups that test-takers could sometimes sound overly direct and forceful and could reject or disagree with their interlocutors in a disaffiliative manner that might be considered impolite. It appears that although both test-takers and their peers agreed that test-takers could possess the linguistic means and control over their emotions, their approach to disaffiliative management might still at times cause strain on social harmony.

6.2.3.2 The affiliation promotion sub-section

For the *affiliation promotion* sub-sections, all six items were within the infit item ranges for both questionnaires, indicating that the two sub-sections measured unidimensional constructs. The infit range for the self questionnaire was 0.43–1.55 while the range for the peer questionnaire was 0.54–1.46. The Rasch test reliability for this sub-section in the self-assessment questionnaire was 0.67 while the one for the peer-assessment questionnaire was 0.75. Both reliability indexes were acceptable for the low-stakes purposes of the questionnaires.

Table 56 presents the fair scores, SE and infit values for the six *affiliation promotion* items in both questionnaires. Similar to the results in the *disaffiliation control* sub-section, test-taker peers on average consistently viewed test-taker performance more positively than test-takers across all six items. Most test-takers agreed that they possessed the ability to use linguistic devices to promote solidarity (item 5) but were less confident about their ability to preserve affiliation at a more holistic level, such as demonstrating enough support and affirmation (item 16) or projecting a caring and concerned persona (item 29). This is understandable as test-takers could have a relatively confident gauge of their use of linguistic resources but might be less certain about the effect of their use of such resources on their interlocutor or the interaction at large. Test-taker peers, on the other hand, held a much more approving view of test-takers' attempt at affiliation management on a global level, giving them on average higher scores for item 16 and item 29 than for the more linguistically-oriented item 5. This shows that test-taker peers in general acknowledged test-takers' attempt at promoting affiliation and viewed their attempt very positively.

Table 56 Affiliation promotion questionnaire item analysis

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
5	Can use language to build rapport	5.17	0.15	1.31	5.20	0.18	0.84
6	Can use appropriate intonation	4.52	0.12	0.73	5.32	0.20	0.83
7	Can demonstrate empathy and sympathy as required in Chinese contexts (RC)	3.47	0.10	1.39	4.72	0.14	1.18
12	Can look after interlocutor's feelings and personal circumstances	4.85	0.13	1.06	5.12	0.18	1.04
16	Can show support and affirmation when conversing	4.79	0.13	0.74	5.35	0.20	0.71
29	Can sound caring and concerned to maintain strong relationships	4.84	0.13	0.69	5.25	0.19	1.39

The item regarding intonation, item 6, attracted contradicting views from the two groups. Test-takers in general were less confident about their ability to use appropriate prosodic features during interaction compared to their opinions on the other items, whereas test-taker peers overall agreed that test-takers could adjust their intonation to the interactional contexts. This highlights that test-takers were less certain about their control over paralinguistic devices compared to linguistic devices, though L1 interactants considered test-takers' command of both equally positively. This suggests that more feedback to test-takers on paralinguistic features, whether in assessment or teaching, will be facilitative to assisting test-takers with developing more confidence in their use of non-lexical interactional features, a point also noted in Dai (2023a).

The two items in relation to empathy and frame (Goffman, 1974; Kim, 2013), items 7 and 12, attracted less positive ratings in both cohorts. This shows that both test-takers and their peers were less certain of test-takers' ability to manage empathetic unions and sharing of frames with interlocutors, pointing to areas worthy of more development and emphasis in teaching and assessment within the IC category *affiliation promotion*.

6.2.3.3 The morality sub-section

The eight items in the *morality* sub-section in both questionnaires also demonstrated acceptable Rasch test reliability, with 0.67 for self-assessment

and 0.79 for peer-assessment. All items were within their respective item infit ranges: 0.53–1.37 for self-assessment and 0.34–1.62 for peer-assessment.

Table 57 Morality questionnaire item analysis

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
10	Can be a collaborative team player in group settings	4.60	0.12	0.82	5.13	0.21	0.79
11	Can avoid appearing casual or not serious enough in Chinese contexts (RC)	3.00	0.10	1.29	5.04	0.18	1.26
13	Can convey one's respect for interlocutors	4.91	0.14	0.94	5.54	0.25	0.96
19	Can appear sincere to effectively de-escalate	4.54	0.12	0.91	5.34	0.24	1.13
25	Can prioritize negotiation and consensus	4.75	0.13	0.83	5.41	0.23	0.84
30	Can appear genuine and sincere in finding solutions	4.70	0.12	0.59	5.22	0.22	0.74
32	Can avoid appearing passive or not proactive enough when there's a problem (RC)	3.68	0.10	0.99	4.75	0.17	1.58
33	Can appear honest to interlocutors	4.99	0.15	1.21	5.55	0.25	0.48

A closer look at Table 57 provides more detail regarding how the items were perceived by test-takers and their peers. Both groups considered test-takers able to demonstrate universal moral qualities such as respectfulness and honesty, as evidenced by the near 5 (agree) responses to items 13 and 33. Test-takers were least confident about their ability to avoid sounding flippant (item 11) and appearing passive (item 32). It is noticeable that most test-takers slightly disagreed with the statement that they could avoid sounding too casual in item 11, with an average self fair score of 3.00 (slightly disagree). Contrary to test-takers' pessimistic view on item 11 and item 32, most test-taker peers gave agreeing responses to items 11 and 32, showing that test-taker peers were comfortable with the degree of seriousness and proactiveness of test-takers. Items 10, 19, 25 and 30 demonstrated similar patterns in score differences as in previous sub-sections in that test-taker peers were consistently giving more positive evaluations than test-takers, with approximately a half-point difference in fair average scores.

6.2.3.4 The reasoning sub-section

The Rasch test reliability indexes for the *reasoning* sub-section were high, with 0.84 for self-assessment questionnaire responses and 0.82 for peer-assessment questionnaire responses. The six items were also within the infit ranges in their respective questionnaires, with 0.29–1.57 for self-assessment and 0.59–1.27 for peer-assessment. This generated confidence in the unidimensionality of the *reasoning* sub-section.

Table 58 Reasoning questionnaire item analysis

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
4	Can follow a clear structure	4.37	0.13	1.54	5.09	0.22	0.84
14	Can provide detailed explanations when needed (RC)	2.10	0.13	1.01	4.01	0.18	1.28
20	Can elaborate ideas and opinions comprehensively	4.20	0.12	0.73	5.11	0.22	0.77
23	Can propose reasonable and sensible solutions when needed	4.73	0.15	0.50	5.17	0.24	0.90
24	Am solution-focused when in difficult interpersonal situations	4.63	0.15	0.91	5.37	0.26	0.79
27	Can provide believable and appropriate reasons when needed	4.80	0.16	0.91	5.39	0.26	0.99

Similar to previous sub-sections, Table 58 shows that test-taker peers demonstrated more leniency in their assessment of test-taker's reasoning ability compared to test-takers' assessment. The rank of fair averages in both questionnaires was in general consistent, with item 27 considered easiest to agree with, followed by items 24 and 23. This indicates that both groups were confident in test-takers' ability to adopt a solution-focused approach to interaction, proposing sensible solutions to overcome interactional difficulties, and providing appropriate reasons when necessary.

Both groups demonstrated slightly less confidence in test-takers' competence in attending to structural issues in interaction, such as providing a clear structure to their talk (item 4) and developing their ideas or arguments (item 20). Item 14 is worth noting in that it was the least positively rated item out of all the 35 items for test-takers. On average test-takers disagreed with the statement that they could provide explanations for their conduct with enough detail, resulting

in a 2.10 fair average for item 14. This shows that test-takers in general were unconfident about their ability to proffer sufficient logical details for their explanations. Test-taker peers, on the other hand, appeared more forgiving in their rating, providing an average 4.01 *slightly agreeing* rating on this item. This again highlights the difference in test-takers' harsher perception of their IC compared to a more relaxed one from test-taker peers.

6.2.3.5 The social role management sub-section

Two separate Rasch analyses were run on the seven items in the *social role management* sub-section in both questionnaires. The infit range for self-assessment was 0.29–1.65. All seven items were within the infit range except for item 28, which was borderline misfitting. Item 28 was kept for this sub-section in the self-assessment because of the unique information it contributes. For peer-assessment, the initial analysis of all seven items showed that item 28 had an infit value of 1.89, which was largely misfitting. After item 28 was removed, the new infit range was 0.44–1.48. Item 26, with an infit of 1.52, was slightly misfitting. Considering the information it contributed to the construct, it was kept in the analysis. The remaining items proved to be reliable measures of the construct, with a 0.80 Rasch test reliability for the *social role management* sub-section in the self-assessment questionnaire, and a 0.75 test reliability for the same sub-section in the peer-assessment questionnaire.

Looking at Table 59, different from the fair averages in previous sub-sections, the gap in score differences between self fair average and peer fair average were much smaller for *social role management*, with less than a half-point difference for most items compared to a one-point difference for items in previous sub-sections. This shows that test-takers and their peers had similarly positive perceptions of test-takers' ability to enact and orient to social roles.

Table 59 Social role management questionnaire item analysis

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
1	Can adjust language to interlocutor's social role	4.94	0.14	0.88	5.36	0.24	0.75
3	Can communicate appropriately based on one's social role	4.86	0.14	0.70	5.34	0.23	0.79
9	Can change one's language when one's social role changes	4.85	0.13	0.66	5.07	0.21	0.80

Table 59 Continued

Item	Content	Self fair average	Self SE	Self infit	Peer fair average	Peer SE	Peer infit
18	Can adapt language to different interlocutor's differing social roles	4.91	0.14	0.85	4.93	0.21	1.00
26	Can demonstrate a good understanding of activities bound to the interlocutor's social role	4.73	0.13	1.26	4.93	0.21	1.52
28	Have a good understanding of Chinese seniority and social ranking (RC)	3.61	0.11	1.66	Misfitting	/	/
34	Can demonstrate strong knowledge of activities bound to one's social role	4.94	0.14	0.76	5.28	0.23	0.87

Test-takers in general considered themselves able to adjust their language to changes in the social roles and concomitant predicates/activities (see Section 2.4.6 for definitions of activities and predicates in MCA) in themselves and their interlocutors, which was evidenced by the score range from 4.7 to 4.9 for items 1, 3, 9, 18, 26 and 34. Test-takers were less confident in their knowledge of the social ranks and seniority systems in Chinese society, giving themselves a 3.61 fair average in item 28, which was around one point lower than the other items.

Test-taker peers, on the other hand, were more positive in test-takers' competence in designing language to their own social roles and matching appropriate predicates to their social roles (items 3 and 34). Peers in general also agreed that test-takers could tailor their language to the social role that the interlocutor possessed (item 1), though they were slightly hedged in their opinion of whether test-takers could match appropriate predicates with their interlocutors' social roles (item 26). Test-taker peers were also less confident in test-takers' ability to handle changes in their own social roles (item 9) and their interlocutors' roles (item 18).

Findings from this sub-section demonstrate that test-takers and their peers in general had consistently positive views of test-takers' competence in managing social roles, though subtle differences exist. Particularly noteworthy is that peers had reservations regarding test-takers' competence in managing changes in social roles, which indexes the ability to adjust language to the specific roles enacted in real life. This observation highlights the challenges for L2 speakers to

effectively adapt their language to the differing social roles both in themselves and in their interlocutors in various interactional settings, which warrants more attention in assessment and pedagogy. Dai and Davey (2022, 2023) and Tai and Dai (2023) are recent empirical studies that used Sequential-Categorial Analysis to uncover the micro-level sequential and categorial resources speakers draw on to shift their social roles. An understanding of how these resources are deployed can assist language learners with developing the social abilities mentioned in this sub-section.

6.2.3.6 Overall results of self and peer IC questionnaires

The sub-sectional analyses conducted above from Section 6.2.3.1 to 6.2.3.5 generated information regarding the relationships between test-takers' self-assessment and their L1-speaker friends' peer-assessment on each of the five IC categories. Now let us examine the functioning of all the items in the self and peer-assessment questionnaires.

For the self-assessment questionnaire, item 35 was identified to be misfitting. If item 35 was removed and a Rasch analysis was rerun on the remaining 34 items, the Rasch test reliability was 0.92, which showed that the self-assessment questionnaire as an instrument was reliable. The infit range for the 34 items was 0.36–1.60 and all items were within the range except for one item at 1.74. This suggests all the items were measuring one unidimensional construct, which was test-takers' self-perception of their IC. The Wright map in Figure 24 presents questionnaire respondents/test-takers, questionnaire items and the questionnaire scale on the same logit scale. The Wright map shows that the questionnaire items were well matched with test-takers' self-perception of their IC.

For the peer-assessment questionnaire, items 28 and 35 were identified as misfitting and were therefore removed. A Rasch analysis of the remaining 33 items generated a Rasch test reliability index of 0.91, which was again a very high reliability index. The infit range for the remaining 33 items was 0.30–1.62 and only one item was slightly misfitting at 1.76. Similar to the self-assessment questionnaire, this shows most of the items in the peer-assessment questionnaire were measuring one unidimensional construct, which was test-taker peers' perception of test-takers' IC. Different from Figure 24, the Wright map in Figure 25 for the peer-assessment questionnaire indicates that most items were too easy for the test-taker peers. This means that test-taker peers were prone to give high ratings on the items, showing that they were inclined to agree with the statements in favour of test-takers' IC. This finding reveals that test-taker peers in general viewed test-takers' IC very positively, which was different from the oft-harsher judgements test-takers formed of their own IC.

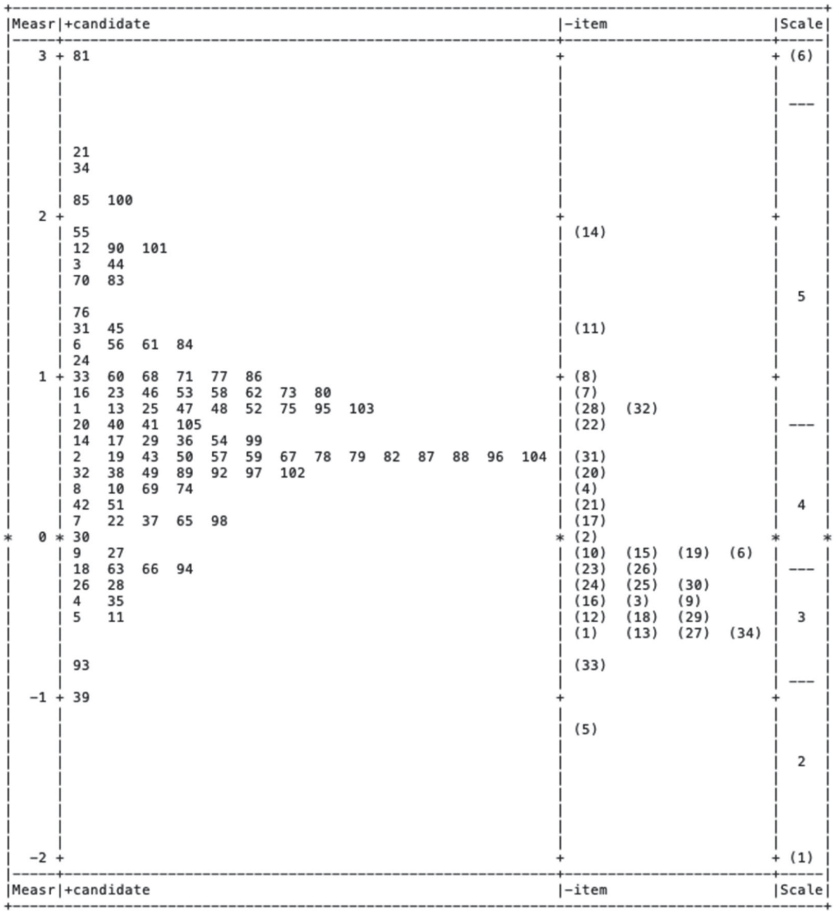


Figure 24 Wright map for self-assessment questionnaire items

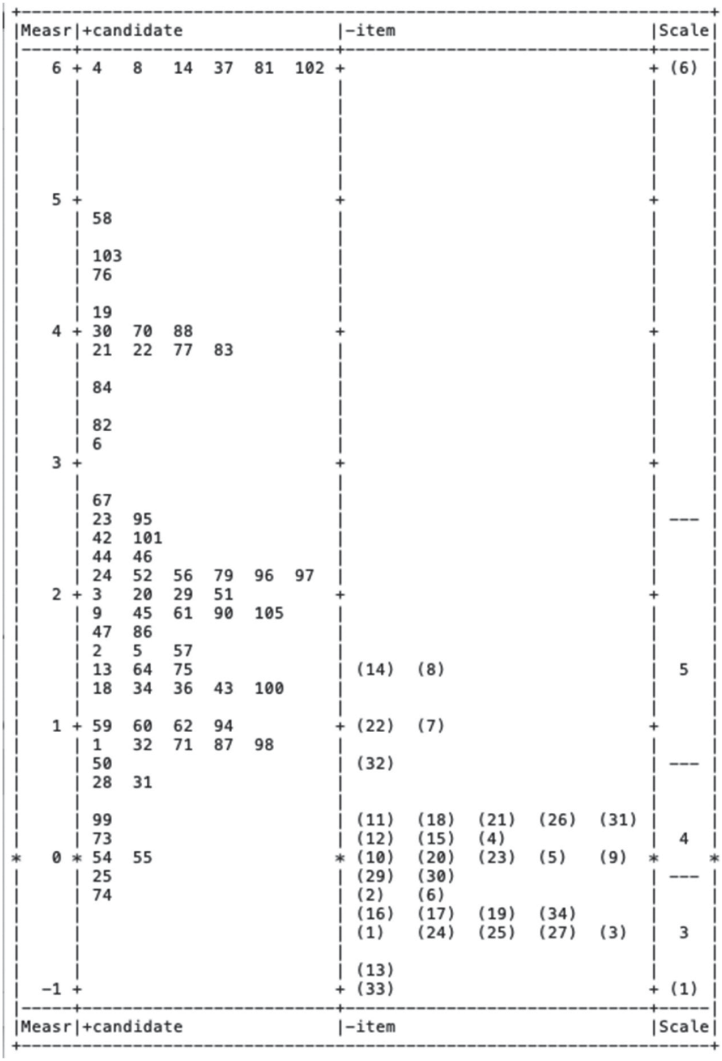


Figure 25 Wright map for peer-assessment questionnaire items

6.2.4 Correlation between the IC test and questionnaires

Having examined the results for the items in each of the five sub-sections in both the self and peer-assessment questionnaires in Section 6.2.3, Section 6.2.4 explores the correlation between questionnaire sub-sections and test-takers' test performance on the five rating categories in the IC tests. The purpose of the correlational analyses was to examine the extent to which test performance could predict test-takers' and their peers' subjective questionnaire-based assessment of test-takers' IC *in the wild*.

Correlational analyses were first conducted on test-takers' Rasch fair scores on each of the five IC rating categories and test-takers' Rasch fair scores on each of the five sub-sections in their self and peer-assessment questionnaires. Since individual Rasch analyses were run for each sub-section in both questionnaires as explained in Section 6.2.3, each test-taker received an average fair score for each of the five sub-sections in their self-assessment questionnaires and a different set of five average fair scores for the five sub-sections in their peer-assessment questionnaires. Separate Rasch analyses were conducted on test-takers' raw scores in each of the five rating categories in the IC tests, generating five fair IC rating category scores for each test-taker. The analyses in this section used the average fair scores from each of the two questionnaires to correlate with the category fair scores from the IC test. Figure 26 explains how the correlational analyses were conducted for each pair. If we use *disaffiliation control* as an example, each test-taker received a Rasch fair score on this rating category in the IC test based on all nine items in the test. If the test-taker completed items related to *disaffiliation control* in their self-assessment questionnaire, they would receive a *disaffiliation control* self fair score via Rasch analysis. If this test-taker's peer completed items related to *disaffiliation control* in the peer-assessment questionnaire, the test-taker would have a *disaffiliation control* peer fair score calculated via Rasch. Correlations were then calculated between test-takers' IC test scores on this particular rating category and their self and peer scores on this category from the two questionnaires.

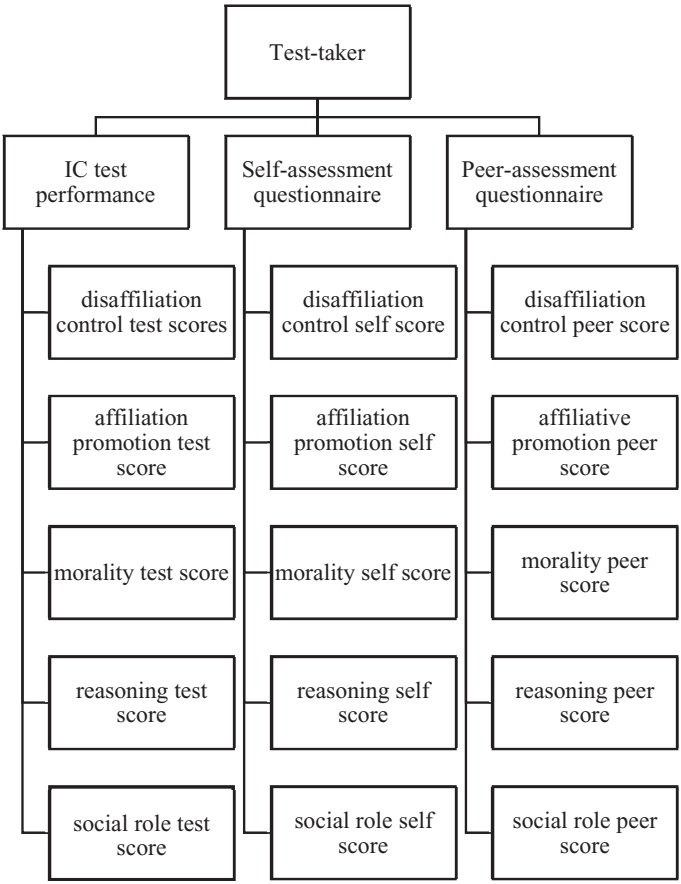


Figure 26 Correlational analyses between the IC test and the questionnaires

Let us first inspect the relationships between test-takers' IC test category scores and their sub-section scores in the self-assessment questionnaire. The results were summarized in Table 60, which shows that there were significant correlations between each of the five pairs with varying strength. Test-takers' average fair score on the seven *disaffiliation control* self-assessment questionnaire items correlated at 0.345 with their fair score in the *disaffiliation control* rating category in the IC test ($N=101$, $p<0.05$). Test-takers' average fair score on the six *affiliation promotion* self-assessment questionnaire items correlated at 0.215 with their fair score in the *affiliation promotion* category in the test ($N=101$, $p<0.05$).

For the *morality* category, test-takers' average fair score on the eight items in this sub-section of the self-assessment questionnaire correlated at a weaker 0.200 with their test fair score on the IC *morality* rating category, but the correlation was nevertheless still significant (N=101, $p<0.05$). The strongest correlation between the sub IC categories was the correlation between test-takers' fair average on the six *reasoning* sub-section items in the self-assessment questionnaire and their fair score on the *reasoning* IC rating category. Pearson Product Moment coefficient was 0.396 and the correlation was significant (N=101, $p<0.05$). Finally, test-takers' average fair scores on the seven items in the *social role management* sub-section in their self-assessment questionnaire correlated moderately at 0.236 with their fair scores on the *social role management* rating category in the test (N=101, $p<0.05$).

A tentative explanation for the stronger correlation in the *reasoning* pair is that evidence for reasoning skills *in the wild* is easier to garner and evaluate compared to the ones for the other four categories. It is putatively less challenging to form an accurate impression via questionnaire items of one's ability at providing solutions and explanations than one's ability at, for example, demonstrating moral characters, promoting affiliation, or enacting social roles.

Table 60 Correlations between test scores and self-assessment scores

	Disaffi test score	**Affili test score**	**Morali test score**	**Reason test score**	**Role test score**
Disaffi self score	0.345*				
Affili self score		0.215*			
Morali self score			0.200*		
Reason self score				0.396*	
Role self score					0.236*

Similar correlational analyses were conducted on the five pairs between the sub-section fair scores in test-taker peer-assessment questionnaires and the IC test category fair scores, which were presented in Table 61. Only the pair between test-taker peers' average fair score for the *reasoning* sub-section in the peer-assessment questionnaire positively correlated with test-takers' IC test scores on *reasoning*. The Pearson coefficient was relatively low at 0.248 (N=73, $p<0.05$). The correlations for the other four pairs were not significant. The results show that test-taker peers' perception of test-takers' IC *in the wild* can only be moderately predicted by test-takers' test performance for the *reasoning* category

alone. Similar to the explanation offered in the self-assessment questionnaire results, one possibility for the significant *reasoning* correlation is that reasoning as an IC dimension is easier to assess through questionnaire items.

Table 61 Correlations between test scores and peer-assessment scores

	Disaffi test score	**Affili test score**	**Morali test score**	**Reason test score**	**Role test score**
Disaffi peer score	-0.490				
Affili peer score		-0.430			
Morali peer score			0.290		
Reason peer score				0.248*	
Role peer score					0.115

Finally, Rasch fair scores were calculated for all the items in the self-assessment questionnaire and peer-assessment questionnaire, generating a fair self-assessment score and a fair peer-assessment score for each test-taker. These two questionnaire scores were then correlated with test-takers' fair scores based on their nine performances in the IC test. The purpose of this analysis was to examine how the self-assessment questionnaire as a whole correlated with the IC test, and how the peer-assessment questionnaire as a whole correlated with the IC test. The findings are presented in Table 62 and it can be observed that there was a moderate correlation between test performance and test-taker self perception at 0.347 ($p<0.05$), while the correlation between test performance and test-taker peer perception was weak and non-significant.

Table 62 Correlations between the test and the two questionnaires

	Total self score	**Total peer score**
Total test score	0.347*	0.056

A few discussion points are in order when we look at the correlations between test scores and self-assessment and peer-assessment scores. In terms of the correlation between self-assessment and the IC test, the correlation is moderate and significant, although it is lower than the ones reported in previous test-criterion correlation studies (Chapelle et al., 2008). A related concern is that

since the self-assessment questionnaire was administered to test-takers after they had completed the IC test, the correlation could have been inflated due to the order effect.

There are three main factors contributing to the lower-than-expected correlation between self-assessment and IC testing. First, the self-assessment questionnaire items were based on DEs' indigenous criteria of IC, instead of the theorized IC test construct. Had the questionnaire items been modelled on the operational IC test construct as represented in the theorized IC rating scale, which was used to score test-taker performance to arrive at the IC test scores, the correlation between self-assessment and the IC test could have been higher. Second, the item statements in the self-assessment questionnaire (see Appendix V) are generic statements pertaining to the indigenous IC criteria, unrelated to the specific test tasks or scenarios in the IC test. If test-takers were asked to self-assess their real-world IC in a questionnaire on the exact nine scenarios covered in the IC test, it is likely that the correlation would be much higher. Third, the concept of IC, especially in the form of the expanded IC model as presented in this book, can be foreign to test-takers in this study. Asking test-takers to self-assess their IC and its related features is very different from asking them to rate their abilities by LC measures such as pronunciation, grammar, or vocabulary, all of which have been consistently and exhaustively explained to L2 speakers in most language classrooms. Therefore, the lower correlation found in this study compared to previous LC test-criterion correlation studies could simply be due to the fact that test-takers were not experienced in providing reliable self-assessment of their IC.

Despite these limitations in terms of alignment between the self-assessment questionnaire and the IC test, the fact that the correlation is still moderate and significant speaks volumes about the extrapolative power of the IC measured in the test to the IC *in the wild*. It shows that test-takers' test performance on each of the five IC rating categories can to a reasonable extent predict their perception of their IC performance outside testing settings.

In terms of peer assessment, although test-taker peers are putatively most knowledgeable of test-takers' IC *in the wild*, the correlations between the IC test and peer assessment, as presented in Tables 61 and 62, are mostly non-significant. The three reasons detailed above for self-assessment can also explain the correlation observed in peer assessment. Test-taker peers assessed test-takers' IC based on impressions from their everyday interaction, which might not encompass or fully cover the task scenarios in the IC test. Test-taker peers did not have access to test-takers' test performance to be able to assess the more specific and targeted IC performances represented in the IC test.

An additional factor that could have affected the peer-assessment results is that the peers selected were test-takers' self-nominated L1 friends. These peers might have a predisposition to assess their respective test-taker friends in a more favourable light due to the existing relationships they had with the test-takers. This is corroborated by the overwhelmingly positive peer-assessment ratings compared to test-takers' self-ratings and test raters' ratings on test-takers' IC test performances. The relationship between test-taker peers and test-takers might have affected the precision of peer-assessment, which then blurred the correlation between peer-assessment and IC testing. Despite all these confounding factors, it is even more noteworthy that the correlation for the *reasoning* pair was significant for peer-assessment. This shows that the results of the IC test could offer a good prediction of test-taker behaviour in everyday contexts when we are looking at the *reasoning* IC category.

It is worth noting that the correlations for both the self-assessment and test pair (0.347) and the peer-assessment and test pair (0.056) are lower than the correlation for the HSK and test pair (0.416). Further analysis can include a hierarchical regression analysis to force in self-assessment and peer-assessment first and then add HSK in the analysis. This can control for the order effect and offer additional evidence for divergent validity between LC and IC.

6.2.5 Rasch analyses of extrapolation and attitude items

Section 6.2.3 and Section 6.2.4 have examined how the rating category-related items in the two questionnaires functioned and how the items correlated with test-takers' test performance. As mentioned in Section 6.1.2.3 in the methodology section of study three, apart from the 35 items targeting the five IC rating categories, the self-assessment questionnaire also contained two additional sections that assessed to what extent test-takers thought their test performance could predict real-world behaviour (six items) and their attitudes towards the IC test (five items). Different from the 35 items in Section 6.2.3 that implicitly assessed test-takers' self-assessment of their real-world IC behaviour according to the five rating categories, the six prediction questions here explicitly assessed how far test-takers thought the test modelled their real-world behaviour and gauged their real-world IC, which elicited direct evidence in support of the extrapolation inference in the validity argument. The five attitude questions, on the other hand, measured test-taker opinions towards the IC test developed. Though test-taker attitude or uptake is not discussed in Kane's validation framework, it is arguably an important measure of a language test's marketability and commercial potential, which is of interest to various end-user groups.

6.2.5.1 Explicit extrapolation questions

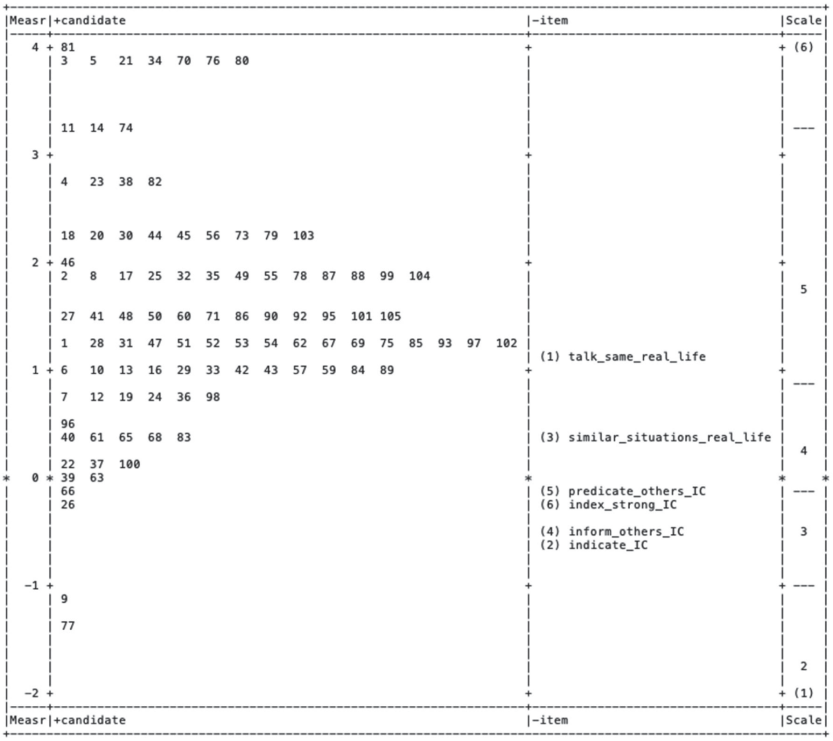


Figure 27 Wright map on the explicit extrapolation questions

Figure 27 offers a visual representation of the matching between the six extrapolation items and the 99 test-takers that completed this section. It can be seen that the items were matched with the lower end of test-takers on the logit scale, indicating that most items were very easy for test-takers to endorse. This shows that test-takers in general were leaning towards agreeing with the statements that supported the extrapolation inference. The Rasch test reliability was 0.70 and all six items were within the item infit range (0.65–1.37), indicating that the items were collectively measuring the same underlying construct.

Table 63 reports the infit value, the fair average score, and their respective SE for each item. Similar to the analyses of other questionnaire items in previous

sections, SEs were reported to indicate the precision of the measurement of the fair average scores.

Table 63 Rasch analyses of the explicit extrapolation items

Item	Content	Infit	Fair average	SE
1	I talk in the same way in real life as in the test scenarios (RC)	0.93	4.08	0.11
3	Test performance is similar to behaviours in similar situations in real life	1.05	4.70	0.12
5	Test can predict people's ability to address interpersonal situations in Chinese	0.75	4.94	0.14
6	High scores on the test indicate strong ability in successful interaction	1.03	4.98	0.14
4	My test scores can inform others of my IC	0.93	5.09	0.15
2	The test says a lot about my interpersonal skills in Chinese	1.36	5.15	0.16

Looking at the Wright map and the item fair scores together, it is clear that test-takers found items 2 and 4 easiest to endorse. Test-takers agreed that their IC test scores revealed their L2-Chinese IC (item 2) and served as a reliable indicator of their IC to stakeholders such as teachers and employers for a range of purposes (item 4). Items 5 and 6 were similar in item difficulty and both of them were slightly harder to endorse compared to items 2 and 4. However, test-takers still mostly agreed that top performances on the test indicated very strong IC (item 6) and the test could reliably predict test-takers' IC in real-world interpersonal situations (item 5).

Items 1 and 3 were the hardest to agree with, though test-takers still adopted an agreeing stance to both statements (*slight agree* was coded 4 and *agree* was coded 5). A fair 4.7 average on item 3 shows that test-takers believed their performance on the test was similar to behaviour in real life in similar situations. This indicates that the tasks in the IC test, the way the tasks were delivered, and the performance elicited through the test tasks all mirror real-world interactional situations and generate reliable information regarding test-takers' IC. Item 1 was the hardest to agree with, with most test-takers adopting a slightly hedged position towards this statement. The fair average of item 1 shows that most test-takers slightly agreed with the proposition that they interacted in real life in

exactly the same way as they do in the test. A more qualified position towards this statement is understandable as assessment can only model reality, instead of fully replicating reality. Testing by its very nature can only elicit samples of real-world behaviour for assessment. Another possible explanation is that item 1 was negatively phrased, which invited test-takers to agree to a more hedged opinion.

Judging from the information the six explicit extrapolation items provided, it can be concluded that test-takers demonstrated strong confidence in the extrapolation of IC test scores to their real-world conduct.

6.2.5.2 Test-taker attitude questions

Figure 28 and Table 64 represent the Rasch findings of the five test-taker attitude questions. The Wright map in Figure 28 illustrates that most items were matched to the lower end of test-takers, indicating that the propositions in the items were very easy for test-takers to endorse. The Rasch test reliability was 0.72 and all five items were within the item infit range (0.54–1.46), suggesting that this five-item sub-section was a reliable instrument to measure a unidimensional construct, which was test-takers' attitude towards the IC test.

Although test-takers displayed an overall agreeing attitude to all five items, a closer inspection of the fair item scores revealed subtle differences in their opinions. Test-takers found item 5 the easiest to agree with. Most test-takers *more than agreed* that Chinese teachers should introduce IC-oriented interactive tasks in the test to language classrooms. This shows that L2-Chinese speakers, as a group of stakeholders in language education and assessment, advocate for a stronger focus on the teaching of interactional skills. Test-takers also affirmed the benefits of the test tasks for developing stronger IC for real-world applications, as demonstrated by the overall *more than agreeing* attitude towards item 1.

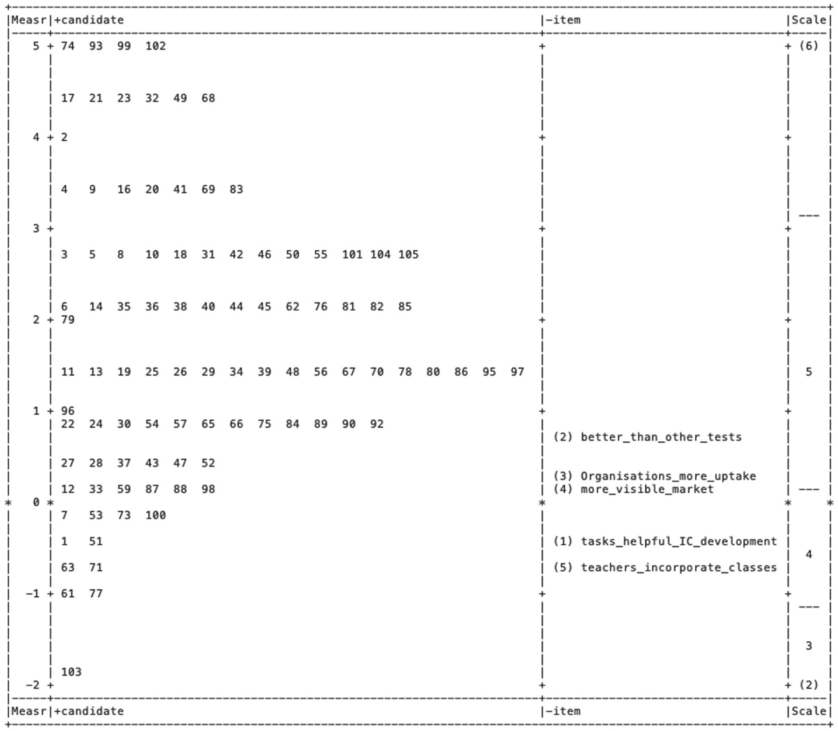


Figure 28 Wright map on the attitude questions

In terms of the commercial prospect of IC assessment, item 4 revealed that test-takers agreed that there should be more similar IC assessment tools or materials on the market. This is encouraging news for test developers and testing companies as we have evidence that IC tests similar to the one developed in this study are welcomed and endorsed by test-takers/end-users. Items 2 and 3 were relatively harder for test-takers to agree with but overall test-takers still maintained a positive attitude towards both statements, with fair averages going well above *slightly agree* (4) and approaching *agree* (5).

Table 64 Rasch analyses of the attitude items

Item	Content	Infit	Fair average	SE
2	This IC test is better than other Chinese speaking tests. (RC)	0.89	4.81	0.15
3	More uptake by organizations to use this IC test to assess IC.	0.81	4.94	0.16
4	More visibility of this IC test on the market.	0.79	5.00	0.16
1	Test tasks can help develop stronger IC.	1.18	5.21	0.18
5	Chinese teachers should introduce test tasks to classrooms.	1.36	5.28	0.18

Item 3 enquired about the uptake by organizations such as universities or companies to use the IC test as a screening tool for language competence. The slightly less agreeing attitude from test-takers could be due to test-takers' differing preferences in how their speaking skills and IC were to be measured in high-stakes settings.

Similarly, a slightly qualified agreeing stance was recorded for item 2, which can be due to test-takers' personal predilections towards assessment tools. It is not unreasonable to expect some test-takers to have a penchant for the assessment tasks that they have been familiarized with from language classes and existing language tests. However, this does not invalidate the usefulness and benefits of IC assessment. As IC assessment and pedagogy become more prevalent, it is reasonable to expect a positive washback and more favourable uptake by test-takers.

Overall, the five test-taker attitude items show that the IC test has met with very positive feedback from test-takers in terms of stakeholder preference, pedagogical benefits, and commercial marketability.

Chapter 7 Validity argument and overall discussions

Chapters 4, 5 and 6 have detailed the three interrelated studies conducted for this book, focusing on test design, rating materials design and the main testing study respectively. If we revisit the purpose of conducting the three studies, we are reminded that the three studies were designed for the purpose of test development and validation, which brings us back to Chapter 3, the *interpretive argument*. In Chapter 3, I introduce Kane's argument-based validation framework as adopted by this book and lay out the inferences investigated in the book, the assumptions behind each of the inferences, and how the gathering of the backings for the assumptions shapes the design of the three studies. Now that the three studies have been presented, Chapter 7 assesses the backings gathered through the three studies and evaluates if they can support the assumptions and inferences in the format of a *validity argument*. In Chapter 7 the assumptions and backings for each of the inferences are investigated individually. For ease of reference, below is a reproduction from Chapter 3 of Figure 1, the validation framework, and Table 6 to Table 8, which explain how the research questions in each of the three studies are built on the assumptions in the validation framework. It should be noted that the decision and consequence inferences (Chapelle, 2020; Knoch & Chapelle, 2017) are not investigated in this book as both inferences are only assessable after the test becomes operational, which goes beyond the scope of this book.

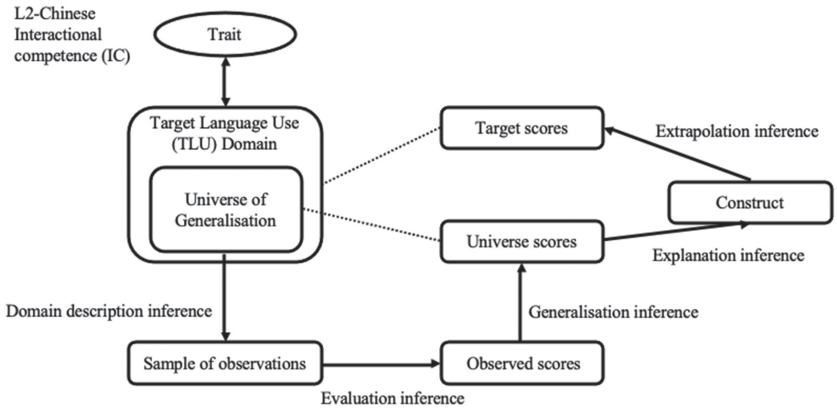


Figure 1 *The validation framework for this book*

Table 6 Research questions for study one

Research questions in study one based on assumptions	
Assumptions	Assumptions reformulated as questions
Domain assumption 1	Can critical activities indexing IC be identified in the TLU domain?
Generalization assumption 1	Are there clear test specifications documents to generate parallel tests?
Domain assumption 2	Can IC assessment tasks be developed that are based on the identified critical IC activities?
Domain assumption 3	Do the IC tasks in the test offer sound coverage of the TLU domain?
Evaluation assumption 1	Are the task administration conditions conducive to the elicitation of IC skills from the IC test?
Domain assumption 4	Do the task scenarios and delivery methods represent activities in the TLU domain?

Table 7 Research questions for study two

Research questions in study two based on assumptions	
Assumptions	Assumptions reformulated as research questions
Explanation assumption 4	Is the IC test construct unidimensional based on pilot testing?
Extrapolation assumption 1	Does the test construct embody the evaluation of critical IC skills in the TLU domain?
Explanation assumption 1	Is the rating scale modelled on the test construct and offering a clear representation of the test construct?
Evaluation assumption 3	Can the steps in the rating scale precisely measure test-takers' different ability levels?
Explanation assumption 2	Is test-taker performance reflective of the test construct?
Explanation assumption 5	Does test-taker IC performance relate to measures of LC/speaking but cover unique IC/talking variance?
Evaluation assumption 2	Are raters well trained to use the rating materials?
Explanation assumption 3	Is the IC test construct in keeping with theories and philosophical assumptions of interpersonal interaction?

Table 8 Research questions for study three

Research questions in study three based on assumptions	
Assumptions	Assumptions reformulated as research questions
Generalization assumption 2	Can the test-takers be sampled to approximate a sound representation of the real-world test-taker population?
Evaluation assumption 2	Are raters well trained to use the rating materials?
Evaluation assumption 3	Can the steps in the rating scale precisely measure test-takers' different ability levels?
Generalization assumption 3	Can tasks in the IC test reliably measure test-taker IC?
Evaluation assumption 4	Do individual raters demonstrate high intra-rater reliability?
Generalization assumption 4	Do different raters demonstrate high inter-rater reliability?
Explanation assumption 4	Is the IC test construct unidimensional?
Explanation assumption 5	Does test-taker IC performance relate to measures of LC/speaking but cover unique IC/talking variance?
Extrapolation assumption 2	Do the IC rating categories measure test-taker IC in similar real-life activities in the TLU domain?
Extrapolation assumption 3	Do stakeholders perceive the IC test as measuring test-takers' ability to interact in the TLU domain?

Chapter 7 also discusses considerations outside the validation framework, including the practicality of assessment, stakeholder attitude towards assessment, stakeholder assessment literacy, the generation of a universal IC model, the applicability of the IC construct and rating scale, and the parameters of the IC tasks.

7.1 The domain description inference

Mirroring the argument structure in the interpretive argument in Chapter 3, let us first inspect the backings for the assumptions for the domain description inference. Table 65 presents the warrant, the assumptions reformulated as questions, the backings for the assumptions and the related studies for the first inference in the validity chain.

Table 65 Domain description inference assumptions and backings

Domain description inference		
Warrant: Test-taker performance on the IC test can reveal test-taker IC in the L2-Chinese TLU domain		
Assumption questions:	**Backings:**	**Studies related:**
Domain assumption 1: Can critical activities indexing IC be identified in the TLU domain?	Disaffiliative social actions in speaking as identified by the task-based needs analysis (TBNA)	Study one
Domain assumption 2: Can IC assessment tasks be developed that are based on the identified critical IC activities?	Tasks modelled on critical incidents, tasks containing rich details that were elicited through the H-S interviews in the TBNA	Study one
Domain assumption 3: Do the IC tasks in the test offer sound coverage of the TLU domain?	Nine test tasks informed by well-sampled TBNA informants with variation in the Power variable, sub-TLU domains and task delivery methods	Study one
Domain assumption 4: Do the task scenarios and delivery methods represent activities in the TLU domain?	Both qualitative and qualitative evidence elicited through four groups of informants, the use of an item panelling committee, and two norming sessions	Study one

As discussed in Section 3.1.1, the domain description inference rests on the warrant that a clearly defined TLU domain can be identified where IC test tasks can be developed out of an infinite pool of similar tasks in the UG within the TLU domain. Four assumptions were identified in order for this warrant to be supported and now we need to examine if sufficient backing was generated for each of the assumptions in the design of the IC test in study one.

7.1.1 Domain description assumption 1

The first assumption requires me to identify critical IC skills and abilities in the TLU domain. This was achieved through a TBNA on the most challenging interactional situations that L2-Chinese speakers frequently experienced. Results of the TBNA indicate that L2-Chinese speakers find launching social actions, especially disaffiliative actions, the most demanding task. The critical incidents the informants reported were mainly focused on the verbal mode of

communication, instead of the written mode. Based on the findings, the critical IC activities identified for this IC test are social situations where L2-Chinese test-takers need to verbally manage disaffiliative social actions.

The process of conducting a TBNA concurrently narrowed the TLU domain. Before the TBNA was conducted, the TLU domain was broadly defined as 'L2-Chinese interaction for L2-Chinese speakers who have a need to live, study and work in China' (see the domain description inference in the interpretive argument in Section 3.1.1). The TBNA refined the TLU domain to 'L2-Chinese verbal interaction in disaffiliative social actions for L2-Chinese speakers who have a need to live, study and work in China'. As discussed in Section 3.1.1, a narrower TLU domain is needed for the design of a language test, though the inference regarding test-taker IC that can be drawn from the test is reduced. In the case of this IC test, it measured test-takers' online productive skills in implementing disaffiliative social actions verbally in three sub-TLU domains. Since the task scenarios in the test are concerned with general, everyday-life interactional incidents in the life, study and work settings, the test is suitable for a wide range of test-takers and allows test users to infer test-takers' ability across a large TLU domain.

In terms of other related skills, the tasks in the IC test to some extent tapped test-takers' sociopragmatic and pragmalinguistic competence, as both types of competence featured in the TBNA results and were incorporated into the design of the tasks (see Section 4.2.2 for how the categories in the TBNA were incorporated in the test). However, if stakeholders are interested in more precise measurements of L2-Chinese speakers' sociopragmatic and pragmalinguistic competence, they need to commission test developers to design more targeted sociopragmatic and pragmalinguistic assessment instruments such as the ones produced in Roever (2005) and Roever et al. (2014) for L2-English. The IC test in this book is also unable to reliably predict test-takers' receptive or written IC skills as all the tasks in the test are delivered in the verbal mode of communication. Despite the limitations in the inferences stakeholders can draw from the results of the IC test, stakeholders can be confident that the IC test is indeed measuring the IC skills that are critical to the target test-taker population and relevant to the stakeholders, judging from the methodical TBNA and the various triangulations the TBNA employed.

7.1.2 Domain description assumption 2

Backing for domain assumption 2 can be found in how results of the TBNA were translated to prototype tasks. The H-S interview technique used in the TBNA

generated accounts of critical incidents with rich details. These critical incidents lent themselves smoothly to IC task scenarios that were authentic and specific to the L2-Chinese context. Though the task prompts in the test simplified the scenarios portrayed in the critical incidents, the detailed descriptions in the task prompts still offered rich contextual and situational information to make the scenarios relatable to test-takers. To make the test tasks represent real-life activities and embody the principle of language tasks in Long (2016), the prompts in the tasks did not prescribe the social actions that test-takers needed to undertake. This made the tasks more open-ended and similar to activities in real life, where L2 speakers have more flexibility and autonomy in deciding what they want to do in particular disaffiliative situations.

7.1.3 Domain description assumption 3

Supporting evidence for domain assumption 3 can be located in a comparison between activities in the TLU domain and the content in the test tasks. The IC test contains nine verbal role-play tasks, which target the eight top-ranking disaffiliative social actions from the TBNA results. The social actions assessed in this test match the specific L2-Chinese context in that the social actions selected for the test are informed by the results of a TBNA in this particular assessment context. This practice has extended the ones in existing IC research where the social actions analysed are either based on researchers' judgement or the available social actions in the data collected. The nine tasks also incorporate a wide range of context-specific social roles, such as doing being a *student*, a *classmate*, a *neighbour*, an *employee*, and a *colleague*. There is also sufficient variation in the power variable, sub-TLU domains (everyday, work and study) and the CMC methods used in the test. In particular, the three CMC methods utilized in this book are aligned with the five CMC-for-IC considerations listed in Section 2.2.3. The three CMC methods (first pair-part voice messaging, second pair-part voice messaging and live video chat) represent a continuum from asynchronous to synchronous, from less interactive to more interactive communication methods. The choice of task delivery methods in this book is therefore not haphazard as it is informed by existing research to increase variability in and coverage of the TLU domain. The CMC platform in the book, WeChat, is a mobile phone-based communication application that is frequently used by L2-Chinese interactants in the TLU domain. The use of WeChat for task delivery makes the test tasks more representative of activities in the TLU domain. It should also be emphasized that the content of the tasks was based on insight from carefully sampled TBNA participants (see Section 4.1.1.1). The various triangulation methods and the

consideration of proficiency and sub-TLU domains employed in the selection of TBNA participants ensured crucial activities across the TLU domain were reported. In summary, though the tasks in a language test can never exhaust the possibilities in a TLU domain or fully replicate the TLU domain, the different types of variation in the nine tasks, the considerations behind the task delivery methods, and the careful selection of TBNA participants who informed the task content have provided support to the argument that the IC test offers sound coverage of the TLU domain.

7.1.4 Domain description assumption 4

Finally, domain assumption 4 is supported by the validation of the tasks through both qualitative and quantitative methods, as detailed in Chapter 4 study one. The involvement of four groups of informants, L1-Chinese speakers, L2-Chinese speakers, language teachers and applied linguists (both pragmaticians and testing specialists) in Section 4.1.1.2, assisted me through multiple iterations of the tasks to make sure the tasks measured the expected test construct and were accessible to the target test-taker population. The pre-pilots and trials of the test tasks with the abovementioned four groups of informants ensured the tasks functioned as expected. The use of an item panelling committee of testing specialists is aligned with the best practice in the field (Alderson et al., 1995). The two rounds of norming questionnaires with L1-Chinese speakers generated evidence that the tasks can elicit the expected responses, the task scenarios are authentic, the match between CMC methods and tasks is appropriate, and the three politeness theory contextual variables function as intended.

In summary, the backings garnered for the domain description inference suggest that a clear TLU domain was defined, critical IC skills and activities were identified and test tasks were based on critical activities in the TLU domain and measured the identified critical IC skills. The tasks offered sufficient coverage of activities in the TLU domain and were validated to be reliable instruments. This supports the movement from the domain description inference to the next inference, the evaluation inference.

7.2 The evaluation inference

Table 66 presents the assumption questions, backings, and related studies for the four identified assumptions behind the evaluation inference. Now let us inspect the backings for each of the assumptions to see if the assumptions are well supported.

Table 66 Evaluation inference assumptions and backings

Evaluation inference		
Warrant: Test-taker performance on the IC test can be scored to generate observed scores that measure critical IC skills.		
Assumption questions:	**Backings:**	**Studies related:**
Evaluation assumption 1: Are the task administration conditions conducive to the elicitation of IC skills from the IC test?	Pre-pilots and trials with experts to decide on prompt comprehensibility, task delivery methods, preparation time and other details in task administration	Study one
Evaluation assumption 2: Are raters well trained to use the rating materials?	The creation of CA-informed test-taker exemplars and the three-step scaffolded approach to rater training	Studies two and three
Evaluation assumption 3: Can the steps in the rating scale precisely measure test-takers' different ability levels?	Engagement with DEs to decide on the number of steps needed. Many-facet Rasch showing adequate person separation and observations on each step	Studies two and three
Evaluation assumption 4: Do individual raters demonstrate high intra-rater reliability?	Rasch rater infit values demonstrating high intra-rater reliability	Study three

7.2.1 Evaluation assumption 1

Validity evidence for evaluation assumption 1 can be sought from an inspection of the task administration conditions. The trials and pre-pilots of the IC test in study one informed me that one of the main challenges in administering the IC tasks was to ensure that the test prompts were comprehensible to test-takers of different proficiency levels. When building test tasks from critical incidents reported in the TBNA, I ensured that the prompts for the test scenarios contained enough contextual details to reconstruct the challenging interactional situations in real life. Richer contextual details also made the IC tasks more challenging for higher proficiency L2 speakers as existing IC assessment tools are not difficult enough for advanced test-takers (Roever, 2018). The trade-off of including rich details to generate difficult IC tasks is that the test prompts can become lengthy, posing processing difficulty to lower proficiency test-takers. In order to not let comprehension become construct-irrelevant variance, I drew on feedback from the four test-design informant groups (L1 speakers, L2 speakers,

language teachers and applied linguists in Section 4.1.1.2) in the trials and pre-pilots in study one and wrote prompts that are comprehensible to test-takers across different HSK levels.

In terms of task delivery, the tasks were presented to test-takers in the video format so that reading and literacy did not influence test-taker performance. Before each video was delivered, I sent the test-takers a list of keywords in the video in both English and Chinese. The words contained names of role-play roles and context-specific words such as *xuemei* 'junior female classmate at school' which might be unfamiliar to certain groups of test-takers. I answered test-takers' questions if they were unclear about the meanings of the keywords. These keywords also appeared in the video prompts in both languages (see the video prompt images in Appendix III). The inclusion of keywords, again, served to ensure that construct-irrelevant variance did not contaminate the measurement of IC.

The use of cartoons in the video prompts contextualized tasks scenarios and aided comprehension. The cartoon character representing the test-taker in the video was deliberately drawn to be gender-neutral so that test-takers of different gender identities could feel comfortable relating to the roles in the prompts (see the video prompt images in Appendix III). The recruitment of a professional news reporter and the controlled speed of speech in the prompts minimized the influence of accent in task delivery and lessened the pressure on test-takers' listening processing capacity.

Drawing on insight gleaned from the pre-pilots and trials, thoughts were also put into the procedures during performance elicitation after the tasks were delivered. The one-minute preparation time was a balance between authenticity and practicality. In real life, test-takers will have some time to think before they initiate a voice messaging conversation, respond to voice messages, or launch a video call with their interlocutor. The time for thinking can vary greatly in real life but needs to be controlled in a testing setting so that the test can be administered in a reasonable time frame. The four groups of informants in trials and pre-pilots informed me that one minute was a suitable length of time for test-takers to prepare and formulate a reasoned response. Test-takers were also given the option to not use up the full one minute if they felt ready before the time was up, which improved the authenticity of the tasks.

The decision to not impose a cap on the number of voice messages a test-taker could send for any particular task was also made in consideration of backings for the evaluation assumption 1. Since in real life there is no limit on the number of voice messages a person can send to address a particular social action, a cap on voice message numbers would have been construct irrelevant. Some test-takers

might prefer to send short 10-second voice messages successively while others might prefer to send one long 55-second voice message for a task. Such variance was tolerated to ensure test-takers could complete the tasks in a manner that was most natural and comfortable to them.

To sum up, the task administration conditions were optimized to ensure the delivery of the tasks could elicit the intended observed performances from the test tasks, without introducing unwanted construct-irrelevant variance.

7.2.2 Evaluation assumption 2

Backings for evaluation 2 in terms of rater training come from both studies 2 and 3. Select pilot test-taker performance excerpts were transcribed and analysed in CA fashion to validate the test construct. Similar to Greer (2020), I subsequently developed test-taker exemplars from these CA-transcribed performance excerpts for the first step in rater training where raters were getting familiarized with the five rating categories.

After raters felt comfortable connecting the rating scale with test-taker performance via the CA-informed exemplars, the ordering activity in the second step in rater training focused on developing raters' ability to spread test-taker performance to different steps in the rating scale. In order to not cognitively overload raters, they were asked to order only one set of performance excerpts against one rating category. As five sets of performances were given, they had the opportunity to practise ordering performances against all five rating categories.

The last step in rater training was a mock rating session where raters were required to rate test-taker performances on all five rating categories. Disagreement in rating was discussed and moderated to ensure standardized perceptions of IC indicators. Correlational analysis was conducted to examine rater agreement before both raters were considered ready for rating.

The three steps in rater training adopted a scaffolded approach to gradually introduce raters to the test construct and the rating scale. During the rater training meetings, raters reported that they felt more aware of what to look for in test-taker performance against the rating scale, after being presented with the CA-informed exemplars. Raters also found the ordering exercise conducive to a more nuanced understanding of the differences in step difficulties. Apart from raters' subjective feedback, the efficacy of rater training was reflected quantitatively in the high intra- and inter-rater reliability indexes from Rasch analyses, which will be discussed soon.

7.2.3 Evaluation assumption 3

Evaluation assumption 3 presupposes that the steps in the rating scale were meaningful to raters and expects that raters used the steps to meaningfully spread test-takers to different ability levels. Backings for this assumption can be found in the rating scale construction phase in study two and the quantitative Rasch analysis in study three. The decision to have a five-step rating scale was based on DEs' feedback that a three-step solution was not adequate to measure the nuanced granularity in test-taker performance (see Section 5.2.3.2). The input from DEs justified the design of five steps in the rating categories as five steps were what DEs needed to differentiate different performances. It needs to be reiterated that in this book both the DEs recruited to inform the test construct and the raters recruited to rate test-taker performances were linguistically untrained (see Section 5.1.1.3 and Section 6.1.1.3). DEs' perceptions of the number of steps needed represent what linguistically untrained raters need to meaningfully differentiate performances. The approach to deciding on the steps in the rating scale differs from the more common researcher-driven *a priori* approach where researchers subjectively decide on the number of steps without consulting raters' feedback (see Section 2.6.1).

In terms of raters' actual use of the rating scale, backings for successful differentiation of test-taker performance can be found in the Wright map, the candidate measurement report, and the scale functioning report in study three in Chapter 6. The Wright map in Figure 21 shows that 105 test-takers were spread across five logits and all five steps were sufficiently utilized by raters. A closer look at the candidate separation index (strata) in the candidate measurement report in Table 46 indicates that test-takers were separated into 8.39 statistically distinct groups. This demonstrates that the rating scale was sensitive enough to distinguish between high IC and low IC test-takers across a wide spread. The small SE of candidate measurement (mean=0.12, SD= 0.01) in Table 46 reveals that test-takers were measured with precision. This is partly due to the fully-crossed rating design adopted but is also indicative of raters' precision in awarding scores. Lastly, the rating scale functioning report in Table 50 provides evidence that there were sufficient observations on each scale step and the rating probability curves in Figure 22 highlight five separate peaks representing each of the five steps in the rating scale. In summary, Rasch analyses generated backings that raters could use the five-step rating scale to separate test-takers across a wide range of abilities, supporting evaluation assumption 3.

7.2.4 Evaluation assumption 4

Finally, the backing for evaluation assumption 4, which is intra-rater reliability, can be found in the Rasch rater measurement report in Table 47 in study three. The mean-square infit values in the rater measurement report are measures of how consistently raters rate. McNamara et al. (2019) argued that high infit values indicate that raters are misfitting in that they are not rating consistently within their own rating. Having analysed the complete data set of test-takers and test-taker performance on all five IC categories and all nine items, the Rasch model could have detected if there were any unusual harshness or leniency in raters' ratings when the model did not expect it. This unexpected randomness could have dampened intra-rater reliability. If the infit values were overfitting, they would be indicators of potential halo or central tendency effects as there was not enough variation in raters' ratings. As discussed in Section 6.2.1.3, the infit values for the two raters, 1.02 and 0.98 respectively, were neither misfitting nor overfitting. This shows that both raters rated in a highly reliable manner within themselves and did not demonstrate undesirable central tendencies or halo effects. This piece of evidence offers sufficient support for evaluation assumption 4.

7.3 The generalization inference

Looking at the validation framework in Figure 1, the movement from observed scores to universe scores requires the generalization inference, which implies that test-taker performance on this particular IC test can be generalized to other parallel IC tests in the universe of generalization (UG). Four assumptions were identified in the interpretive argument and now let us examine the assumptions and their respective backings, as presented in Table 67.

Table 67 Generalization inference assumptions and backings

Generalization inference		
Warrant: Observed scores are reflective of expected scores across parallel tasks in the UG.		
Assumption questions:	**Backings:**	**Studies related:**
Generalization assumption 1: Are there clear test specifications documents to generate parallel tests?	Sufficient information provided in the test specifications. Test tasks and delivery methods standardized to facilitate item writing	Study one
Generalization assumption 2: Can the test-takers be sampled to approximate a sound representation of the real-world test-taker population?	Purposeful selection of test-takers from different L1 backgrounds, age groups and professions	Study three
Generalization assumption 3: Can tasks in the IC test reliably measure test-taker IC?	High Rasch test and item reliability indexes and low SEs	Study three
Generalization assumption 4: Do different raters demonstrate high inter-rater reliability?	High rater agreement, similar rater logit measures, low rater separation and strata	Study three

7.3.1 Generalization assumption 1

Generalization assumption 1 expects me to develop clear test specifications of the IC test so that parallel IC tests can be generated. Backing for this assumption is provided in Section 4.2.2 in study one when I translated TBNA results to IC test tasks. The presentation of the test specifications in Table 12 followed the checklist in Alderson et al. (1995), which covers most of the specifications information needed for different groups of stakeholders. In particular, the structuring of the IC test adopted the three contextual variables – power, social distance, and rank of imposition – in politeness theory. The use of the contextual variables assisted with the standardization of the broader context where social actions took place so that item writers in the future can generate parallel tasks from the UG. The three task delivery methods – 1st pair-part voice messaging, 2nd pair-part voice messaging and video chat – are also standardized CMC delivery methods that are easy to replicate when writing new items. Other information such as the intended test-taker population, the rating scale and the test construct was also defined clearly in different parts of the book. To sum up, the details of the

specifications of the IC test are sufficiently provided for item writers to generate parallel versions of the IC test, which supports the generalization inference.

7.3.2 Generalization assumption 2

Backing for the second generalization assumption can be located in the selection of test-takers in study three. Roever et al. (2014) acknowledged that researchers in most applied linguistics and language testing studies tend to recruit tertiary students, which only represent one group of L2 speakers and limit the range of generalization to the broader target L2 speaker population. This issue was addressed in this book via the accessibility brought by the CMC assessment platform. The IC test was designed to be delivered entirely online, making it possible for the researcher, who was based in Australia, to recruit test-takers from different countries and continents around the world.

Apart from the ease of test-taker recruitment due to the CMC nature of the test, potential test-takers were screened to offer a sound representation of the target L2-Chinese speaker population. Out of the 90 L2-Chinese test-takers, 30 spoke a variety of English as their L1, 30 spoke an Asian language as their L1, and 30 spoke a non-English, non-Asian language as their L1. The recruitment of relatively more English-L1 test-takers was due to ease of access as I resided in Australia, speaks English and had easier access to L1-English test-takers. It should be noted, however, that a range of L1-Englishes was included to improve generalizability, such as Australian English, American English, British English, and Nigerian English. For the 30 Asian-language-L1 test-takers, they were mainly from L1 backgrounds such as Vietnamese, Thai, and Korean. The selection of these test-takers offered a better representation of the target test-taker population and avoided an overrepresentation of test-takers from Indo-European L1 backgrounds. The remaining 30 test-takers came from language backgrounds such as Italian, Urdu, German, Russian, Farsi and Hebrew, which represented a wide range of language families. The diverse linguistic backgrounds of test-takers (21 different L1 languages and 7 bilingual combinations, see Section 6.1.1.1 for details) strengthened the generalization inference through the sampling of a wider and more representative group of test-takers.

Apart from linguistic diversity, the 90 L2-Chinese speakers came from different walks of life, including undergraduate students, post-graduate students, L2 speakers working in Chinese-speaking companies in China, academics working in universities, L2 speakers working in China-related industries in their home countries, and naturalistic L2 speakers who were strong in speaking but had limited L2-Chinese literacy. The test-takers also represented a diverse

range of age groups, ranging from below 20 to above 40. More details about the test-takers can be found in Section 6.1.1.1. It is worth restating that the reason this book was able to sample from a wider test-taker population is due to the CMC testing platform adopted, which broke down the barriers in traditional F2F assessment. This allowed for the test scores to be generalized to a larger group of potential test-takers and made language assessment more accessible, especially at times when global pandemics (e.g., the COVID-19 pandemic) made F2F assessment largely suspended.

7.3.3 Generalization assumption 3

As discussed in the interpretive argument in Section 3.1.3 in Chapter 3, the reliability indexes for generalization assumption 3 selected in this study are the Rasch test reliability, Rasch item reliability, and SEs. The first two reliability indexes relate to the reproducibility of test-taker and item measures. In other words, they are concerned with whether the test-takers and test items that are measured to be more competent and more difficult are indeed more competent test-takers and more difficult items, if the test-takers in this IC test took parallel versions of the test, or if the items were given to comparable groups of test-takers. These two reliability indexes can estimate how reliably one can generalize the results of this one single IC test administered to this one single group of 105 test-takers to an infinite number of parallel IC tests in the UG and an infinite number of test-taker groups in the entire test-taker population.

In terms of criteria of test reliability, high-stakes language tests such as TOEFL and IELTS have test reliabilities above 0.9 (Roever et al., 2014). Item reliability, on the other hand, is not traditionally reported in analyses using classical testing theory but is an important reliability estimate as it pertains to how reliable item difficulties are when the test is given to different groups of test-takers from the same target test-taker population. As reported in Section 6.2.1 in study three, both the Rasch test reliability and Rasch item reliability for this IC test were 0.97. These are very high reliabilities considering that Rasch underestimates reliabilities compared to traditional measures such as Cronbach's alphas (Linacre, 1997). It shows the IC test has the potential to be used in high-stakes settings as its results can be reliably generalized to parallel tests and different groups of test-takers.

Compared to the two Rasch reliability indexes, SEs provide a different measure of reliability. As discussed in the interpretive argument, SEs are concerned with the precision of measurement. This is an important piece of reliability evidence as the scores from the IC test need to be shown to be able to be generalized to other

test-takers, items, and raters with precision. The average SEs in the candidate measurement report, item measurement report and rater measurement report were 0.12 (SD=0.01) for test-takers, 0.04 (SD=0.00) for items, and 0.02 (SD=0.00) for raters. This shows that the IC test had stable and precise estimates of test-taker ability, item difficulty, and rater consistency, bolstering the trustworthiness and reliability of the test results. Bond and Fox (2015) specified that the desirable low SEs can be achieved when the test items are well matched with the test-taker sample, the steps in the rating scale are more limited, and the test-taker sample size is large. Considering that there are five steps in the rating scale and the test-taker sample in this study is only 105, the low SEs in this study indicate that the test can reliably and precisely measure test-taker abilities.

7.3.4 Generalization assumption 4

The last assumption for the generalization inference relates to the reliability of the ratings between raters. As discussed in the interpretive argument, rater reliability is an important consideration in rater-mediated assessment as raters need to be able to assess test-taker performance in a comparable manner. Test reliability is threatened if a test-taker performance is considered a band 5 performance by one rater, but a band 2 performance by another rater. In view of this, inter-rater reliability was measured by two reliability indexes, rater agreement and comparable leniency and severity, in this book.

For rater agreement, the Rasch rater measurement report in Table 47 shows that out of 4725 opportunities for inter-rater agreement, the two raters reached exact agreement 2208 times, which was 46.7 % of the total ratings. To the best of my knowledge, there is no benchmark for desirable exact rater agreement for rater-mediated language assessment using analytic rating scales. However, intuitively, the exact agreement index in study three is relatively high, considering that each of the five rating categories contains five steps for the raters to choose from. The two raters in this test agreed on the awarding of a score on a particular step in a particular rating category for nearly half of the cases, which shows a very high degree of agreement.

The other inter-rater reliability index, the one indexing the similarity between raters in terms of leniency and severity, can be found in the measures of raters in the logit scale and rater separation index in Table 47. As discussed in Section 6.2.1.3, there was only a 0.04 logit difference between the two raters' measures, showing that the two raters in study three were very similar in how lenient or severe they were when awarding scores. The 'fixed all-same' chi-squared test in Table 47 tested the hypothesis that the two raters rated in the same manner in

terms of severity. Since the p-value for this test was 0.05, which was not significant, it indicates that the two raters were not rating in ways that were significantly different from each other. This is rare in language assessment as human raters usually demonstrate some variation in their rating (McNamara et al., 2019). The low separation (0.93) and strata (1.58) in Table 47 further confirmed that the difference in severity between the two raters' ratings was minimal. Finally, the rater reliability index was 0.46, which indicates there was no reliable difference in the rater measures. In other words, the Rasch programme was not confident of the rating difference between the two raters, which demonstrates high inter-rater reliability.

Taking into account the exact rater agreement and the separation measures in the Rasch rater measurement report, we can be confident that the raters for the main testing study had high consistency between them. This can be the outcome of the clear descriptors in the rating scale and the scaffolded approach to rater training. High inter-rater reliability indicates that we can be confident that test-taker performances will be scored in similar manners when they are rated by different raters, providing backing to the generalization inference.

7.4 The explanation inference

The explanation inference in the validity argument requires evidence supporting the warrant that expected scores in the UG are reflective of the underlying test construct in the TLU domain. This inference is crucial to the validation of a test construct, especially when the construct is a language construct, which is in general difficult to define. As explained in Section 3.1.4 in the interpretive argument in Chapter 3, the task of undergirding the explanation inference in this book was made more challenging because the test construct was developed bottom-up from DEs' indigenous criteria. Now let us inspect the backings gathered for the five assumptions supporting this inference to see if the backings are sufficient to justify the assumptions. The assumptions and backings and related studies are presented in Table 68.

Table 68 Explanation inference assumptions and backings

Explanation inference		
Warrant: Expected scores reflect an IC test construct in the TLU domain.		
Assumption questions:	**Backings:**	**Studies related:**
Explanation assumption 1: Is the rating scale modelled on the test construct and offering a clear representation of the test construct?	Three rounds of thematic analyses to decide on the rating categories in the rating scale, the number of steps in each rating category, and the sub-categories within each step	Study two
Explanation assumption 2: Is test-taker performance reflective of the test construct?	Sequential-Categorial Analysis using CA and MCA on select test-taker performance	Study two
Explanation assumption 3: Is the IC test construct in keeping with theories and philosophical assumptions of interpersonal interaction?	The transformation of an indigenous scale to a theorized scale via CA and MCA concepts. The theorized scale and the embodied IC construct are congruent with the Hymesean conceptualization of communicative competence and the Aristotelian artistic proofs of effective persuasion	Study two
Explanation assumption 4: Is the IC test construct unidimensional?	Unidimensionality evidence from Many-facet item and rating category infit statistics from both the pilot test and the main test. No additional dimensions on the rating categories from Rasch residuals PCA	Studies two and three
Explanation assumption 5: Does test-taker IC performance relate to measures of LC/speaking but cover unique IC/talking variance?	Qualitative backing from Sequential-Categorial Analysis. Quantitative backing from the differentiation between IC and LC in the pilot test and main test. LC conceptualized as a vehicle for IC in rating scale construction	Studies two and three

7.4.1 Explanation assumption 1

Explanation assumption 1 expects the operational rating scale to be a clear representation of the test construct. This assumption interrogates the rigorousness of the rating scale construction process. Study 2 in Chapter 5 details how the rating scale was developed, showing that the final five rating categories and the sub-categories within the categories were all empirically motivated and grounded.

After thematically analysing DEs' comments and extracting their indigenous criteria,I collapsed some of the criteria and selected the top five criteria as the five rating categories (Section 5.2.2 and Section 5.2.3.1). I subsequently conducted a second round of thematic analysis to sort DEs' comments into five steps within each of the five rating categories (Section 5.2.3.2). A third thematic analysis was run within each step in each rating category to decide on the sub-categories and sub-category descriptors within each of the five categories (Section 5.2.3.3). This process resulted in three to four sub-categories for each rating category. Due to this methodical approach, there was a consistent number of sub-categories within each step in each rating category. For example, the sub-category *linguistic devices* within the rating category *disaffiliation control* appeared in each of the five steps in that rating category. This level of high consistency is a product of the three rounds of systematic thematic analyses adopted. Analysing DEs' comments into sub-categories also facilitated the writing of scale descriptors. The band descriptors in the indigenous rating scale were not based on my intuition or training, but on DEs' language from their every-life lived experiences. The resulting indigenous IC scale, therefore, captured DEs' indigenous criteria and offers sound coverage of the test construct, with minimal 'construct shrinkage' (Knoch et al., 2020, p. 1).

7.4.2 Explanation assumption 2

Explanation assumption 2 expects the test-taker discourse to be reflective of the test construct in the rating scale. This was validated in Chapter 5 study two where I conducted Sequential-Categorial Analysis, combing CA and MCA, of test-taker performance to examine if the interaction between test-takers and their interlocutors at the micro level mirrored DEs' comments and the five rating categories. This process connected the discourse-external etic DE perspective with the discourse-internal emic participant perspective. A Sequential-Categorial (CA and MCA) inspection on the interaction between test-takers and interlocutors enriched, validated, and theoretically expanded DEs' atheoretical comments and their indigenous criteria. More details of this process can be found in Section 5.2.4.

7.4.3 Explanation assumption 3

Explanation assumption 3 assumes that the test construct measured in this IC test is consistent with existing theories of interpersonal interaction. This assumption needs to be well supported in that it would be counterintuitive if a test construct cannot be meaningfully related to any existing theoretical or operational models

of interaction. This endeavour is even more necessary for test constructs that are developed from indigenous criteria, which as Brindley (1998) and Shohamy (1996) rightly critiqued, are atheoretical in nature. To address this concern, I drew on DEs' comments and the Sequential-Categorial Analysis of test-takers' performance data to theorize the indigenous rating scale into a theorized IC rating scale in study two Chapter 5.

The resultant theorized five rating categories in Section 5.2.5 offered a broader conceptualization of IC that is not covered in existing IC assessment research. The five categories, *disaffiliation control*, *affiliation promotion*, *morality*, *reasoning* and *social role management,* have moved beyond the mechanistic focus on sequential IC markers (May et al., 2020) and ventured into the emotional, moral, logical, and categorial dimensions of interaction. The sequential aspect of interaction is not overlooked in this IC construct either. When we look at the sub-categories in the final theorized rating scale in Section 5.2.5, we can see that sequential concerns, such as *turn-taking*, *attentive listening* and *preference structure*, were captured in the descriptions of *action formation*, *intersubjectivity* and *social solidarity*. In other words, the mechanistic aspect of interaction was not neglected in the IC test construct in this book but was assessed from the perspective of maintaining social solidarity, preserving the moral order of interaction, and enacting appropriate social roles.

A higher level of theorization also ensured the five rating categories in this book can be applied to other IC assessment tasks and contexts. Most interactional contexts require varying forms of emotional engagement with the interlocutors, logical presentations of one's argument, projection of positive moral characters, enactment of one's social roles, and orientation to the roles of the interlocutors. It is difficult to envisage a form of interaction where no emotional engagement is needed or where logic is disregarded. The moral aspect of interaction is endemic to any form of interaction as the maintenance of social order and interactional order is essentially the maintenance of moral order. The same principle applies to social role management. The sociocultural and categorial aspects of interaction cannot be stripped from interaction as test-takers, whether in general proficiency testing or specific purpose testing, are always talking in specific roles, such as *students*, *classmates*, *nurses*, *pilots*, and *employees* to their interlocutors who are *lecturers*, *fellow classmates*, *patients*, *air traffic controllers*, and *managers*. Therefore, whether it is language assessment for academic discourse, healthcare profession registrations, aviation language, or business interaction, we cannot escape social roles and the culture and context where such roles are embedded.

From a cross-disciplinary perspective, the universality of the five IC categories in this book is attested by its presence in theories of interpersonal communication

in other academic fields such as sociolinguistics and philosophy. Both the Hymesean communicative competence and Aristotelean artistic proofs encompass similar concepts of interaction as captured in the IC test construct in this book. This corroborates the theoretical vitality and robustness of the definition of IC in this book and promises wider applicability outside the immediate IC assessment context where this IC test is positioned. This theoretical vibrancy improves the practicality of IC assessment. The rating scales and rating materials in existing IC testing research are well targeted to their local assessment contexts and test tasks, which improves the precision of measurement but reduces the applicability of the rating scale. It also raises concerns about the cost and practicality of IC assessment as it is expensive to analyse test-taker discourse for every specific language task and to develop context-fitting IC rating criteria that cannot be easily adapted to other contexts. Due to the highly abstract and theory-conforming nature of the rating categories in this IC test, the five IC categories can serve as general rating indicators in other assessment contexts. A trade-off of a highly theorized rating scale and descriptors is that when applied to local assessment contexts, test designers need to adapt the rating scale so that it can be interpreted in the specific assessment tasks and contexts in their respective IC tests. In the current version of the theorized rating scale in Appendix IV, the sub-categories and the descriptors are highly theoretical and abstract. Different sub-categories may have differing importance and weighting for other IC test tasks and contexts. Test developers are advised to localize and modify the content in this rating scale against the specific purposes, needs and features of their IC tests.

7.4.4 Explanation assumption 4

The fourth assumption pertains to the unidimensionality assumption of Rasch analyses. A dimension in Rasch is the latent trait that is being measured by Rasch, and in the context of language assessment, this is the construct a test claims to measure. Rasch requires that only one dimension is measured at a time so the unidimensionality assumption must hold for Rasch analyses to be valid (Bond & Fox, 2015; McNamara et al., 2019). In the context of this IC test, unidimensionality implies that the test is measuring one single ability or construct, which is a test-taker's IC.

The inspection of unidimensionality traditionally follows two steps. The first is to look at the fit statistics as erratic fit statistics are suggestive of subdimensions in the data. The second is to conduct a Rasch residuals PCA to look at if there are meaningful subdimensions in the residuals after the main dimension has been extracted by Rasch. In this book, evidence for unidimensionality was first

gleaned from the pilot test on the 22 pilot test-takers in study two Chapter 5. As reported in Section 5.2.1, all nine items were within the acceptable infit range.

Similarly, in the main testing dataset consisting of 105 test-takers' performances in study three Chapter 6, the infit statistics for the items, test-takers and rating categories were all within the expected infit ranges. A closer look at the point-measure biserial correlations for the rating categories and items in Table 48 and Table 49 show that all items and rating categories had positive correlations. This indicates that the functioning of the items and rating categories conformed with the underlying primary dimension extracted by the Rasch model. There was no suggestion of the existence of multiple dimensions.

To further interrogate the dimensionality of the dataset, a Rasch residuals PCA was conducted on the five rating categories in Section 6.2.1.7. No meaningful dimensions emerged from the residuals, proving that *disaffiliation control*, *affiliation promotion*, *morality*, *reasoning* and *social role management* could all be arranged on a single measurement dimension, which was IC. Considering the evidence from fit statistics and Rasch residuals PCA, it is clear that test-takers were being measured in the same terms, which means the items and the rating categories were measuring the same latent construct. This also suggests that it is meaningful to add a test-taker's scores on different items or different rating categories to generate a sum score of their IC on the whole test.

In summary, the enquiry into the dimensionality of the test construct demonstrates that although the five rating categories tap different aspects of IC, they as a whole assess the same test construct, which is test-takers' ability to handle interaction. The nine items in the test have also proven to measure the same construct.

7.4.5 Explanation assumption 5

The last assumption, explanation assumption 5, relates to the divergent validity of IC assessment. If the IC test in this study claims to measure an IC construct, it needs to demonstrate that it is indeed measuring IC, instead of other related constructs such as linguistic competence (LC). From a commercial perspective, although a strong argument can be made in regard to using primary instead of secondary indicators to measure IC (Scriven, 1987), testing companies and relevant stakeholders are unlikely to abandon speaking tests that have already been developed in favour of IC tests unless the latter can be shown to cover variance not already accounted for by the former. Both quantitative and qualitative backings were sought to support this assumption.

In terms of qualitative backings, the Sequential-Categorial Analysis of test-taker performances in Section 5.2.4 point to the separation between LC and IC. When we look at the performances from Xiaoxin and Eric on the same item, though Xiaoxin's LC was much higher than Eric's, Xiaoxin was perceived to be weaker in IC than Eric by DEs. Sequential-Categorial Analysis of their performances revealed how Xiaoxin mishandled *disaffiliation control* and *social role management*, despite his excellent control of vocabulary and grammar. In stark contrast, though Eric's talk displayed linguistic infelicities, his employment of *affiliation promotion*, appeal to the *moral* order of interaction, and deployment of common-sense *reasoning* ensured successful interaction with his interlocutor. The well-coordinated manner that the interaction between Eric and his interlocutor unfolded offers discourse-internal evidence of Eric's high IC. Other excerpts in Section 5.2.4, such as the performances from Brian and Hans, also support the argument that L2-Chinese speakers, despite their limitations in LC, can mobilize their linguistic devices to demonstrate high IC. This concurs with existing literature such as Lee and Hellermann (2014) and Jenks and Brandt (2013) where lower-proficiency L2 speakers still managed to conduct interaction successfully, drawing on knowledge of social roles and prosodic devices.

In regard to quantitative evidence, it can be found both in the pilot study and the main testing study. Table 26 in Section 5.2.1 ranks the 22 pilot study participants according to their Rasch scores. Even based on the three pilot raters' intuitive understanding of IC we can see that L1 speakers did not invariably outperform L2 speakers. This shows that when being rated on the criterion of interactional success, L1 speakers do not always have an insurmountable advantage over L2 speakers simply because of their strong LC. The disconnect between IC and LC becomes more pronounced after study three was conducted and the 105 test-taker performances were rated against the IC rating scale. Similar to Ockey et al. (2015), HSK, the more linguistically-oriented LC measure, had very limited predictive power on test-takers' performances on the IC test. The correlation between test-takers' HSK scores and their IC test scores was appreciably lower than the ones in existing IC assessment studies where researchers also correlated their test-takers' IC scores with proficiency measures (Ikeda, 2017; Youn, 2013). The explanations for this difference can be because previous studies used actual speaking tests such as TOEFL or IELTS to determine test-takers' LC, whereas the HSK used in this study did not explicitly assess test-takers' speaking ability but assessed their general LC. The use of HSK was due to practical considerations as there are no large-scale standardized speaking tests that are currently in use by L2-Chinese test-takers. Future research can look into if the correlation changes

when the IC test results are correlated with a test measure that specifically taps test-takers' LC in speaking.

Another possible explanation for the lower correlation in this book is caused by the IC construct measured in this test. Both Ikeda (2017) and Youn (2013) based their IC constructs on inspecting CA-transcribed discourse from test-takers who were pre-grouped by LC measures. This introduced a higher overlap between IC and LC as their measures of IC were to some extent predicated on test-takers' LC. The IC test construct in this book, however, was arrived at from the rater perspective, or more specifically, everyday-life DEs' perspective. The design of the instrument that measured test-taker IC in this book was therefore independent of the LC measure HSK. The resultant IC rating scale in Appendix IV, which represents the IC test construct, targets IC indicators that depart noticeably from traditional measures of LC. This different approach to the differentiation of IC and LC likely contributed to the smaller overlap between IC and LC in this book.

When we look more closely into the relationship between IC and LC in Section 6.2.2 in study three, we can gain a deeper understanding of how LC interacted with and mediated IC. As shown in Table 53 and Table 54, no matter how test-takers were grouped based on their LC/HSK, the variation in their IC grew as their LC increased. This shows that LC was a prerequisite for strong IC performance in the test but did not guarantee high IC. IC, therefore, needs to be assessed independently of LC.

Going back to the theoretical discussion in Sections 2.3.4 and 2.3.5, the dataset from this IC test reveals there were test-takers with strong LC but weak IC, such as some of the test-takers in the HSK5, HSK6 and L1-speaker groups. However, there were fewer test-takers with low LC but high IC. This could be because raters only relied on the audio recordings of test-taker performances to make judgements of their IC, which did not allow for the rich paralinguistic communication as noted in Levinson (2006) that might have facilitated interaction. It is possible that if test-takers' body language were taken into account and conceptualized into the IC construct, some lower LC test-takers could have been assigned higher IC scores due to successful employments of nonverbal, multimodal resources. Another possibility is that the nine items in the test were all very difficult so no matter how resourceful low-LC test-takers were, they still lacked the necessary, foundational LC to achieve high IC.

A final point to make regarding the relationship between LC and IC in this test is that there is a degree of LC in the IC test construct. If we recall how the test construct was developed from DEs' indigenous criteria in Sections 5.2.2 and 5.2.3, DEs' original 12 indigenous criteria did contain three more linguistically

oriented criteria: *linguistic choices*, *prosodic features*, and *the structure of talk* (see Table 27). However, these three criteria were mentioned less often by the DEs and were not treated as the be-all-and-end-all criteria of effective interaction by DEs. Instead, they were perceived as vehicles for achieving interactional success and were only considered relevant for the functional purposes they served. In view of this, the three LC criteria were subsumed under the indigenous IC criteria *conflict management*, *solidarity promotion* and *reasoning skills* (see Table 28). After the process of theorization and looking at the three theorized IC rating categories, *disaffiliative control*, *affiliation promotion* and *reasoning* in the IC rating scale in Section 5.2.5, we can see there were still explicit mentions of linguistic, paralinguistic and cohesive devices in the scale descriptors (see descriptors in Table 36, Table 37 and Table 39). However, these LC descriptors were similarly not used as assessment criteria in and of themselves. Instead, they were interpreted as how test-takers used linguistic, paralinguistic, and cohesive devices to launch recognizable social actions and achieve interactional goals. This shows that both conceptually and operationally, the IC test was designed to measure test-takers' ability to handle real-world interaction with all the real-world implications, instead of producing speech that was merely linguistically sound.

In summary, both the qualitative and quantitative backings in the discourse analyses, correlational analyses, and rating scale construction support the assumption that the IC test measures IC instead of LC.

7.5 The extrapolation inference

The final inference investigated in this book is the extrapolation inference, which stipulates that the universe score covering the UG not only points to a clearly defined test construct but can also be extrapolated to a domain score that measures activities in the TLU domain. Three assumptions were identified in the interpretive argument in Section 3.1.5 and now let us examine if the backings generated from the three studies can support the three assumptions in the validity argument. Table 69 lays out the assumptions, backings, and related studies for the extrapolation inference.

Table 69 Extrapolation inference assumptions and backings

Extrapolation inference		
Warrant: the IC test construct reflects the quality of test-taker interaction in the TLU domain.		
Assumption questions:	**Backings:**	**Studies related:**
Extrapolation assumption 1: Does the test construct embody the evaluation of critical IC skills in the TLU domain?	Basing the test construct on the indigenous criteria from a large number of DEs. The resultant test construct representing the critical skills valued by interactants in the TLU domain. Meaningful connections between band levels and TLU domain implications established	Study two
Extrapolation assumption 2: Do the IC rating categories measure test-taker IC in similar activities in the TLU domain?	Significant correlations in all five sub-sections in the self-assessment questionnaire with the five IC rating category scores. Significant correlation between the *reasoning* sub-section in the peer-assessment questionnaire and the *reasoning* IC category score	Study three
Extrapolation assumption 3: Do stakeholders perceive the IC test as measuring test-takers' ability to interact in the TLU domain?	Questionnaire items with explicit propositions showing that test-takers found the test a reliable measure of their ability and activities in the TLU domain	Study three

7.5.1 Extrapolation assumption 1

Extrapolation assumption 1 requires that the test construct be reflective of the criteria used in the TLU domain. This assumption was shored up by the use of DEs' indigenous criteria in the specification of the test construct. There are three main advantages associated with the recruitment of DEs in relation to extrapolation assumption 1. The first is that the use of DEs made the test construct more representative of the assessment criteria used in the TLU domain (Knoch & Chapelle, 2017). The 36 DEs selected for study two in Chapter 5 represented the everyday-life members of society who are more frequently the ultimate judges of L2 speakers' IC than linguists or language teachers (Sato & McNamara, 2019). A test construct that was built on everyday-life DEs' judgement is more reflective of the evaluation of abilities in the TLU domain than a construct based on my intuition or other pre-existing models that might not fit the TLU domain of the test in question. The use of a large number of DEs ensured the resultant indigenous criteria approximated the general consensus of social members in

the target community, instead of representing a few members' potentially skewed opinions.

The second advantage of using DEs' indigenous criteria to develop the IC test construct is that the resultant rating categories in the rating scale represented IC indicators that members in the TLU domain valued the most. As discussed in Section 2.5.1 in the literature review, one of the challenges of IC assessment is the plethora of potential IC indicators that could be included in a rating scale, which was well captured by the various leaves and branches indexing IC indicators in the IC tree in Galaczi and Taylor (2018). The final five rating categories for this IC test measured higher-order IC indicators from the sequential, moral, logical, emotional and categorial dimensions. Table 28 shows the five categories covered most of the comments DEs made about all nine items. Considering the wide range of scenarios the test encompassed in the TLU domain, it is safe to assume the five rating categories did represent what everyday-life interactants in real life actually prioritize in the TLU domain when the focus is on interactional success. The use of DEs has broadened our existing understanding of IC as everyday-life DEs assisted me to move beyond the established sequential focus of IC assessment to the categorial, moral, logical and emotional dimensions of interaction, which were considered crucial by real-life interactants in the TLU domain.

The third advantage of using DEs in terms of extrapolation assumption 1 is that DEs' feedback helped create meaningful connections between band levels in the rating scale and real-world implications. As Table 29 highlights, the five bands in the rating scale correspond to how test-takers will fare in real life and the type and degree of training they need in order to function well interactionally in real life, based on DEs' judgement. This strengthens the extrapolation inference as test results are made relevant to activities and consequences in the TLU domain.

7.5.2 Extrapolation assumption 2

Extrapolation assumption 2 requires backings that can show that the five rating categories in the rating scale measured similar activities in the TLU domain. One classic method to generate backings for this assumption is to elicit perceptions of test-takers' ability in the TLU domain and examine the correlation between perception and test performance (Chapelle et al., 2008; Roever et al., 2014). If reasonable correlations obtain, there is reason to believe that the rating scale is measuring abilities in activities that are found in the TLU domain.

In study three both self-assessment and peer-assessment were employed to gauge to what extent test-takers' test performance reflected abilities that were

found in similar activities in real life. To ensure optimal comprehension by the test-takers and test-taker peers, the questionnaire items were written in plain language and described activities commonly found in the TLU domain. Section 6.2.4 shows that there were moderate correlations between all the five IC rating category scores with the five corresponding sub-sections in the self-assessment questionnaire. The correlation was also established between the self-assessment questionnaire as an entirety and the total IC test score. The significant correlations between all five pairs in the self-assessment questionnaire highlight that the IC test and the rating scale measure abilities in similar activities in the TLU domain. One noteworthy feature of the questionnaire design is that both questionnaires were divided into five sub-sections corresponding to the five rating categories, with each questionnaire sub-section containing multiple items targeting a rating category at different levels. This level of detail in the depiction of the activities in the TLU domain in the questionnaires provides greater confidence that the questionnaires offer a comprehensive representation of the TLU domain.

In the peer-assessment questionnaire, only the *reasoning* category score in the test correlated significantly with the *reasoning* sub-section in the questionnaire. The peer-assessment questionnaire as a whole did not correlate significantly with the total IC test score either. This could be related to the unreliable nature of peer-assessment and the fact that the peers in study three did not base their estimate of test-taker ability on test-takers' actual test performance. More detailed explanations are offered in Section 6.2.4.

In summary, the significant correlations found between the self-assessment questionnaire and the IC test attest to the argument that the IC test and the accompanying rating scale measured abilities in activities found in the TLU domain. A more refined approach to peer assessment that utilizes other observation methods might have generated more informative findings in terms of the connection between test performance and real-world performance as perceived by peers.

7.5.3 Extrapolation assumption 3

Extrapolation assumption 3 needs backings that can strengthen the argument that end-users of the IC test perceive the test as measuring real-world abilities that the test claims to measure. Different from the detailed questionnaire items that implicitly targeted the five rating categories as required for extrapolation assumption 2, the questionnaire items for this assumption explicitly asked test-takers about their perception of this test vis-à-vis their real-world conduct. Findings reported in Section 6.2.5.1 show that test-takers in general found

the test a good measure of their real-world behaviours and an informative assessment tool of their IC *in the wild*. The explicit knowledge from test-takers was informative because test-takers had first-hand experience of the test and could reliably assess how the test compared with their lived experiences in the TLU domain. These questionnaire items also garnered important information for test commercialization as test-takers as stakeholders represent a powerful voice in the testing industry. A test that is found to be an unreliable measure of real-world behaviour by test-takers is unlikely to gain traction in the business.

7.6 Considerations outside the validity framework

Section 7.1 to Section 7.5 have examined the backings for the assumptions under each of the inferences investigated in the validity argument of the book. In this section, four additional considerations are discussed that are outside the Kanean argument-based framework.

7.6.1 CMC and practicality

The first consideration is the practicality of testing. Roever et al. (2014) critiqued Kane's validation framework for its neglect of the practicality of testing and maintained that practicality should be one of the primary concerns of language assessment. No matter how valid a test is or how many backings a test designer can garner for the test's validity argument, an impractical test is unlikely to be used in the real world. If testing companies are unable to administer the test in a cost-effective manner or if the cost of sitting a test is too high for the majority of test-takers, there will be limited take-up of the test in large-scale testing settings. Therefore, practicality is the main criterion and a crucial enabler for real-world test use, the lack of attention to which is a noticeable shortcoming in most existing language test validation frameworks. As practicality does not fit easily in any of the inferences in the Kanean validation framework, the practicality of the IC test in this book is examined separately.

Since the inception of the IC test in this book, practicality has been a central concern in test design and development. The use of CMC as a platform for test delivery is central to enhancing the practicality of IC assessment. The literature review of CMC in Section 2.2 in Chapter 2 emphasizes the vital role of CMC in contemporary L2 learning and teaching. The sudden onset of the COVID-19 global pandemic has made the language learning and testing community more cognizant of the challenges and opportunities of CMC, since during the COVID era F2F assessment was suspended for most parts of the world (Isbell &

Kremmel, 2020; Ockey, 2021). Developing CMC-based assessment is therefore an imperative and the IC test in this book is aligned with the needs of stakeholders and the market.

Second, the design of CMC-based tasks in this book is not an *ad hoc* afterthought but was foregrounded in the test design. The unexpected start of the COVID-19 pandemic meant that many language tests had to move from paper-based or F2F-based to CMC-based in a very short period of time. Some of the tasks in these tests, originally designed for F2F settings, might not transition smoothly to the online space. Different from this *ad hoc* approach, the tasks in the IC test in this book were designed from the outset with CMC built into them. The task scenarios for all the nine items situated interaction in CMC settings. The task validation process, both qualitatively and quantitatively in Section 4.2.3 and 4.2.4, gathered backings for the authenticity of the task scenarios and the appropriateness of using CMC methods in these scenarios. In other words, the IC test was designed specifically for CMC and validated purposefully in terms of whether the CMC scenarios were authentic and whether the CMC methods adopted were reasonable.

Third, the IC test in this book accentuated the practicality and additional benefits of CMC assessment. During data collection for the main testing phase, I attracted a wide range of test-takers of various educational and professional backgrounds from 26 different countries, more than half of which are ODA recipients (OECD, 2021). This level of variability in the test-taker population would be difficult to achieve for this book if the test were conducted in F2F settings. The easy access of CMC ensured that different test-taker groups could participate in this test, despite geographic distance, schedule differences, and socioeconomic constraints. The wide range of test-takers in this book attests to the role CMC plays in democratizing language assessment by making it accessible and affordable to traditionally underrepresented test-taker groups. The roleplay nature of this test could have been resource intensive to administer for testing organizations and expensive to undertake for test-takers in F2FC conditions. The CMC platform adopted in this test has therefore lowered the cost of test administration, improving the practicality of the test.

Fourth, the holistic IC construct developed in this study represents an attempt to address the inherent resource-intensive nature of IC assessment. As IC focuses on micro-level analysis of interaction, real-world implementation of IC assessment can be costly and impractical if test developers need to identify IC markers for their specific IC task and their particular L2 language. This book has developed a holistic IC construct focusing on five separate IC dimensions and a corresponding theorized IC rating scale that has the potential to be applicable

to a wide range of IC tasks and target languages. Future IC test developers can use the resources developed in this book to examine if they can be adjusted to fit their local assessment context, without having to generate new IC constructs and rating materials from bottom-up micro analysis of interactional data. This could potentially improve the practicality of IC assessment.

A final concern related to test practicality is the analytic IC rating scale developed in this book (see Appendix IV). Due to the CA and MCA parlance employed in the scale descriptors, training raters to use the scale might require some additional unpacking of CA and MCA terminology, especially for professional language testing raters who are more versed in the language of traditional proficiency-based rating scales. However, as Section 6.1.3.2 explains, the CA-naïve raters recruited in this book did not find the comprehension of the IC rating scale challenging, once I had explained the concepts of CA in layperson language. It is reasonable to expect that both naïve and professional raters would find the IC scales in the book accessible after targeted training is provided. Although it implicates that rater trainers for this IC scale will need to possess a certain degree of CA and MCA knowledge, this prerequisite is understandable if speaking assessment is to become more IC-oriented. Standardized rater training materials and workshops based on the training process in Section 6.1.3.2 can also be developed to facilitate rater training, addressing the practicality concern.

7.6.2 Stakeholder take-up and assessment literacy

Another consideration that does not fit easily into the validation framework is stakeholder perception and take-up of the test. Existing test validation research focuses greatly on substantiating the claim that the test is measuring what it is supposed to measure, without paying much attention to how the test is received by end-users. This book addressed this gap by eliciting the attitude of one group of stakeholders, test-takers, highlighting how they perceived the usability of the IC test. Using Rasch to analyse the questionnaire items which measured test-taker attitudes towards the test, Section 6.2.5.2 in study three shows that test-takers considered the test more targeted than existing L2-Chinese tests on the market, confirmed the efficacy of the test for learning, encouraged greater take-up of the test by educational institutions and language teachers, and advocated for greater visibility of IC assessment on the market. This is encouraging news for test developers as it shows that end-users welcome CMC-based role-play tests that tap L2 speakers' IC.

The issue of stakeholder take-up is also related to the development of language assessment literacy for test users, a topic that has received increasing attention in

the field of language testing (Kremmel & Harding, 2020; Tsagari, 2020; Coombe et al., 2020). Section 6.2.5.2 discusses the finding that the test-takers in this study were little aware of how existing L2-Chinese speaking tests underrepresent interactional abilities and how L2 IC assessment tasks, such as the ones in the IC test in this book, could have offered a better reflection of their interactional/talking abilities. This points to a gap in stakeholder assessment literacy in that all test user groups, such as test-takers, test designers, language teachers, test development companies, and test-using institutions, need to be made aware of alternative assessment instruments so that they can make informed decisions regarding which language tests to endorse. A lack of understanding of assessment instruments and what each instrument measures can have knock-on effects and implications: stakeholders can misinterpret test scores because they have assumed the scores to reflect abilities that the tests were not designed to measure or cannot measure.

In summary, developing stakeholders' assessment literacy of L2 IC can improve their understanding of the need for assessment to have interactional components like the ones in this IC test. When stakeholders realize the importance of assessing IC, it will generate positive changes in speaking assessment and test designers will be encouraged to include more interactional items in their test batteries.

7.6.3 Building a universal model of IC

This book has shown so far that test development and validation is a lengthy, expensive, and resource-intensive process. A valid question readers might pose at this point is that after all the work detailed in previous sections and chapters, how the findings from this book can be applied to test developers who do not work with the immediate context in which this book situates, such as L2 Chinese as the target language or role plays as their task types. In other words, what broader implications for language education and assessment can this book offer? Sections 7.6.3 and 7.6.4 set out to sketch the theory-generating nature of the IC construct in this book and the applicability of the construct outside its immediate context.

An exploratory undertaking this book took on was to build a universal model of IC. The unified IC model goes beyond the specificities of this book: L2-Chinese, certain disaffiliative social actions, the three sub-TLU domains and the role-play task format. The IC construct in this book has the potential to be applied to a wider variety of L2s, social interaction contexts, language use contexts and types of interaction. This venture was motivated by the context-fittedness of IC constructs and rating scales in existing IC assessment research.

As reviewed in Section 2.6 and various other places throughout the book, the influence of CA on the conceptualization and analysis of IC has continuously encouraged assessment researchers to utilize test-taker performance data from specific language tasks to generate test constructs and rating scales (Galaczi, 2013; Ikeda, 2021; Kley, 2019; Youn, 2015). The resultant constructs and rating scales fit closely to their particular target languages and task formats, have precision in the measurement IC for their specific test-taker cohorts, but lack generalizability to other languages and tasks. This makes IC assessment a costly enterprise that stands in sharp contrast to traditional psycholinguistically-driven LC assessment where LC indicators such as fluency, pronunciation, accuracy, and grammar cut across different languages and task formats.

Drawing on a range of sources to generate the rating scale (Knoch et al., 2021), this book has built an IC construct that transcends its immediate target language, task format and TLU domain. The use of everyday-life DEs to generate indigenous criteria, the theorization of indigenous scale through Sequential-Categorial Analysis (using CA and MCA) of test-taker performance data, and the positioning of the IC construct in broader sociological and philosophical models on interpersonal communication have ensured the theoretical robustness of the IC construct in this book (a detailed discussion on this point can be found in the explanation assumption 3 in Section 7.4.3). Since the emotional, logical, moral, and categorial dimensions of interaction are integral to any form of human interaction, the IC model specified in Figure 19 and Figure 20 (reproduced below) is in theory of universal relevance and applicability, regardless of the target language, language use domain, or type of interaction.

This clearly is a daring claim, and rebuttals can be launched immediately when we look to interactive speaking tasks such as jigsaw tasks and picture description tasks, where there appears to be little to no evidence of use of *affiliation promotion*, projection of *morality*, or *enactment of social roles*. The seeming divide between socially-oriented IC tasks such as role-plays and less socially-oriented IC tasks such as discussion tasks can be reconciled if we position language tasks in their natural, sociocultural contexts that are rich in emotional, logical, moral and categorial fingerprints.

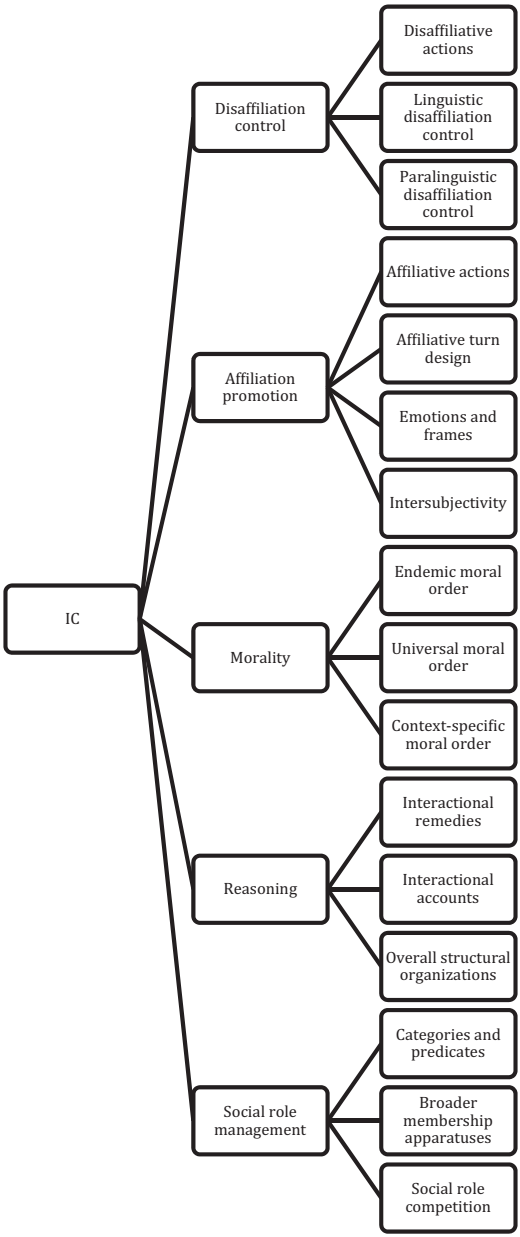


Figure 29 (reproduced) A schematic representation of the IC test construct

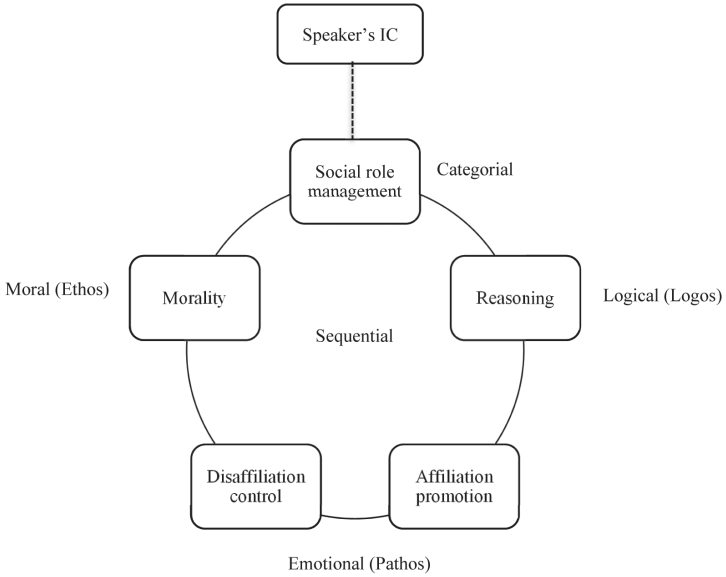


Figure 30 (reproduced) Theoretical import of the five IC categories

Let us take the discussion tasks such as the ones in Cambridge English as an example. If we first revisit the psycholinguistic, LC-driven perspective on language tasks, we see indicators of lexical resources, grammatical range and prosodic features, all of which are measured devoid of the social and interactive functions LC resources fulfil in real-world interaction. Existing IC assessment challenges the psycholinguistic model and advocates for the assessment of IC features such as turn-taking, topic development, topic shift and attentive listening (Galaczi, 2013). These IC features, nevertheless, are only indexing the mechanistic, sequential aspect of interaction (May et al., 2020), without sufficiently accounting for the emotional, logical, moral and categorial aspects of interpersonal interaction. If we take a socially oriented lens to inspect the same discussion task, we notice how these additional aspects are perennially in existence. Even though the discussion tasks might have been designed to foreground the sequential nature of interaction without consideration of sociocultural factors, if we position such tasks in real-life contexts, we quickly see the other dimensions of interaction at play. When a discussant with high IC engages in a discussion with another person in real life, no matter what the discussion topic is, they will be keen to build rapport (affiliation promotion),

to manage moments of disagreement (disaffiliation control), to project themselves as friendly and easy to talk to (morality), to maintain logic when they speak (reasoning), and to sound role-appropriate, depending on whether they are talking as a friend to a friend, or a student to their lecturer (social role management). This shows that there is a social orientation to speaking tasks even when the tasks were not primarily designed to elicit the social dimension of interaction, the reason for which is that human interaction is essentially a social endeavour.

This social nature of interaction is inherent also in monological tasks such as giving a presentation on a designated topic. In real life a monological speaker is always giving a speech to an audience, never to a vacuum. If fact, the Aristotelian artistic proofs were initially conceived for orators who needed to persuade their audience through the appeal to the logical, the emotional and the moral dimensions of communication. Therefore, the IC construct designed in this book is theoretically applicable to any form of speaking assessment as long as the speaking tasks are conceptualized in an interactive, socially oriented manner. The IC model should be operationalizable regardless of whether the task format is monologic or dialogic, whether the task types are roleplays, interviews, discussions, or jigsaw tasks, whether the target language is Chinese, English, or Arabic, and whether the TLU domain is in the clinical, academic, everyday-life, or legal settings.

A point worth emphasizing is that the sequential aspect of interaction is not discounted in the present IC construct (also see the discussion in Sections 5.2.5 and 5.2.6). In fact, the realization of the emotional, moral, logical, and categorial dimensions of IC is predicated on the sequential, turn-by-turn unfolding of interaction. The difference between the IC construct this book proposed and previous mechanistically oriented IC constructs (May et al., 2020) is that the sequential properties of interaction – such as turn-taking, maintenance of intersubjectivity, and topic development – are embedded in the broader emotional, moral, logical, and categorial dimensions of interaction for the current IC construct. This is evidenced by the sequential descriptors in the theorized rating scale in Section 5.2.5.

Building on this IC construct derived from the context of language assessment and to further delineate the reflexive co-determining relationship between sequence and categorization, I have developed a revised model of IC in Figure 31. In this model a speaker draws on sequential and categorial resources to conduct interaction at the emotional, logical and moral levels. Speakers use sequential resources to structure the temporal aspect of interaction, which is synonymous with the concept of time; with categorial resources, speakers talk

social roles/identities, interpersonal relationships, group memberships, entities, institutions and social structures into existence, which is captured by the concept of space. Just like in physics change in space affects time (and vice versa), the same can be said about speakers' use of sequence and categorization. When a speaker for example wants to borrow a large sum of money from someone, depending on the person (categorial) to which they make the request, they adapt their interactional style in terms of for example how many preliminary moves are considered necessary (sequential).

Speakers use sequential and categorial resources to construct a spacetime lifeworld for the main business of interaction: functional language use (Kramsch, 1986). Mirroring the classic tripartite model in communication and rhetoric theories, here I define the main functions of language use/interaction to be to engage with the interlocutor at the emotional, logical and moral dimensions. To interact is to connect, to engage and to aim to approximate intersubjectivity.

This revised IC model takes off from the empirical work in this language assessment project but further specifies the relationships among the different assessment categories in the IC test construct in Figure 29 and Figure 30. Following this second-level theorization, the IC model in Figure 31 offers an improved representation of the various constituents of IC, specifies the relationship between interactional resources (sequential and categorial) and interactional goals (emotional, logical and moral), revisits the functional roots of IC and situates IC in the classic tradition of communication theory. Interested readers can refer to Dai (2023c, 2024) and Dai and Davey (2023) for further discussion of these points.

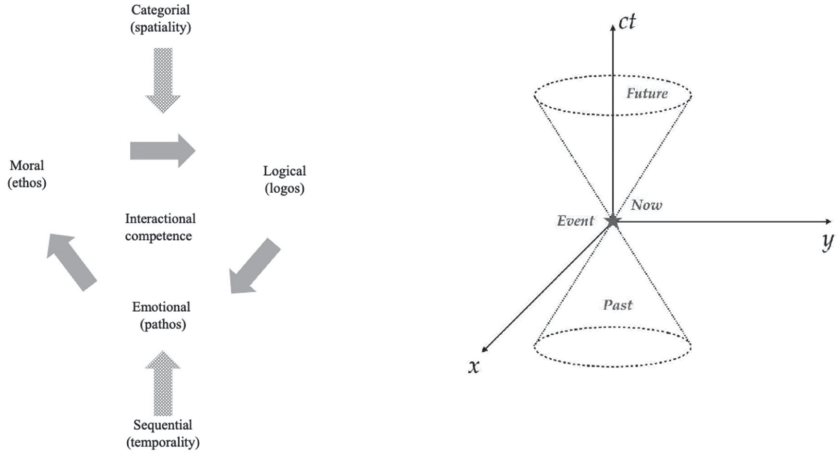


Figure 31 Revised IC model

7.6.4 Application of the IC construct and rating scale

Section 7.6.3 explains the universal IC construct this book has attempted to build but the discussion of it is largely speculative at this point. After a theoretical model has been proposed, the next step is for other test developers to examine it, apply it to their local context, and evaluate if the model holds. Section 7.6.4 offers a brief discussion of the considerations for future researchers if they were to apply the IC construct and the theorized IC rating scale from this book to their respective assessment projects.

The first step in adapting and localizing the present IC construct and scale is to understand the theoretical underpinning behind the dimensions of interaction conceived in this book: namely the emotional, the logical, the moral and the categorial, facilitated by the sequential nature of interaction. The sub-categories and descriptors in the theorized IC scale in Section 5.2.5 were deliberately written in a neutral manner that is not specific to any language, culture, task type, or language use domain. Although the context-independent nature of the theorized IC scale can make the scale appear abstract, this is necessary for a theoretical scale. When the theoretical scale is adapted and localized to a specific language, culture, task type and TLU domain, the scale developer can modify the language in the scale to make it more user-friendly for scale users. In many ways the rating categories in this IC scale are not dissimilar to traditional LC rating categories such as grammar, pronunciation, and accuracy, which could

also appear esoteric to readers unfamiliar with linguistic parlance. The five IC categories developed in this book represent a social, interactional take on interaction, while LC categories embody a psycholinguistic, individualist take on interaction. Both types of categories and category descriptors are abstract and apply to different languages, cultures, and contexts.

When applying the theoretical IC construct and scale, test developers may need to draw on different sources to inform their local IC construct and scale. It is possible that the five rating categories and their sub-categories are of differing relevance to their local context. Some sub-categories might need to be omitted or rewritten to suit the needs of the particular language, TLU domain or task type. The number of steps may also need to be increased or decreased depending on the test purpose and test use context. It is likely that for a language task that assesses a test-taker's ability to order coffee it is not possible or necessary to separate test-taker abilities into five levels on the *morality* criterion. Perhaps the differentiation between a *polite*, *average,* or *impolite* customer is all that is needed by score users for an L2-English IC low-stakes classroom assessment task where the students need to role-play ordering coffee in a café. This process can similarly benefit from the DEs in the local TLU domain as this book did, apart from the many other sources that inform a test construct (Knoch et al., 2021). It is crucial to let the local parameters and test use contexts empirically guide the localization of a highly theoretical and abstract IC construct/scale.

7.6.5 The parameters of the IC tasks

An interesting finding from this book, which is related to future applications of the current IC construct in terms of task design, is that there are no noticeable patterns of influence on item difficulty from the three task parameters: the power variable in politeness theory, the sub-TLU domain, and the degree of synchronism in task delivery methods. Although I varied the nine items in the test by the three aforementioned variables to achieve comprehensive coverage of the TLU domain (see Table 19), the item measurement report in Table 49 shows that all items were within 0.70 logit of difficulty.

Explanations for this phenomenon are manifold. Research in L2 pragmatics assessment has not generated conclusive evidence on whether there is a correlation between the power variable and item difficulty. Neither did Brown and Levinson make any claims in their politeness theory that the larger the power difference between the interlocutors, the harder it is to accomplish the social actions. The influence of the power difference on test-taker production is more likely to be on how test-takers format their social actions, such as the amount of

preliminaries used, the degree of development of the topic, or the proffering of accounts for their action. It is understandable that lower proficiency test-takers may struggle with the complex language expected in items with a high power difference, but once test-takers have attained a certain degree of LC, such items may no longer pose a challenge.

The lack of influence of the sub-TLU domain on item difficulty can be due to the transferrable nature of IC across language use contexts. Although test-takers needed to launch social actions in three different sub-TLU domains, the interactional methods they had to mobilize were likely to be similar, irrespective of whether it was talking in a workplace context or a tertiary education context.

The degree of interactiveness of the tasks, from the least synchronous 1st pair-part voice messaging to the most synchronous live video chat, did not generate any clear difference in terms of item difficulty either. This can be due to the complex, multifaceted nature of the IC construct measured in the test. Existing research on a/synchronous CMC research, as reviewed in Section 2.2, predominantly focused on surface-level morphosyntactic features and controlled for the processing time participants had under different conditions. The degree of interactiveness of the tasks might have less impact on test-taker performance in this test since IC is a more holistic construct that does not focus solely on certain linguistic features. The fact that test-takers had sufficient time to prepare their answers, a condition that is expected in real life, could also have mitigated the item difficulty patterns noticed in more controlled, experimental conditions.

Research in the future can conduct more systematic and refined discourse-level analysis of test-taker performance to see if there are any consistent patterns in test-taker interaction which are influenced by the three parameters. Future IC test development projects can also apply the current unified IC model and rating scale and adjust the three parameters to see if findings from this book still hold.

Chapter 8 Conclusions

Chapter 8, the last chapter of this book, offers a review of this research project, highlighting its significance, innovation, and limitations. Future research direction is discussed.

8.1 Significance of this book

Although this book is concerned with test design and validation, it first started with a philosophical discussion on the most prominent schools of pragmatics theories that inform our current understanding of interaction. This discussion is necessary as there is epistemological tension between the practice of assessment and the analytical policy of Conversation Analysis (CA) on interaction (Roever, 2018). Instead of rigidly following a particular school of thinking, this book proposed a unified model of interaction that reconciled the philosophical and epistemological conflicts in theories, generating an operational definition of interaction in the assessment context in Section 2.1.5. This made the assessment of IC theoretically possible.

Having contributed to the philosophical debate behind IC assessment, the second contribution this book makes is by situating its investigation of IC in a less-commonly taught language on a less commonly used testing platform. The choice of Chinese complements the existing research base of IC assessment, which privileges L2-English. The employment of computer-mediated communication (CMC) as a testing platform not only enriches our existing knowledge of how interaction unfolds in CMC contexts but also responds to the need of our current time, considering how much language teaching and testing have shifted to the online space since COVID-19.The IC test designed in this book shows that IC assessment in the online space is operationalizable, practical, cost-effective, and enthusiastically taken up by stakeholders.

Third, this book adopted a methodical approach in its use of the Kanean argument-based validation framework. Though it is not novel for language assessment research to take an argument-based approach to test validation (Chapelle et al., 2008; Ikeda, 2017; Roever et al., 2014; Youn, 2013), this book went one step further in that it clearly spelled out the assumptions for each of the inferences in the interpretive argument and reformulated the assumptions to research questions to guide the design and execution of the three studies in the book. Table 1 to Table 5 explain the assumptions behind the inferences in order

for the inferences to be valid in the interpretive argument. The assumptions were then translated into research questions, which informed the design of the three studies from Table 6 to Table 8. Once the three studies were completed, backings gathered from the studies were evaluated based on how well they supported the assumptions in the validity argument from Table 65 to Table 69. This systematic approach ensured that the arguments in test validation were carefully evaluated and that necessary precautions were taken in the design of the book to generate the required evidence to make the test a valid instrument.

Fourth, I took rigorous steps in designing and validating the test tasks. The nine items in the test were based on a TBNA on the most challenging interactional situations for L2-Chinese speakers. A range of triangulation techniques was employed to ensure the task-based needs analysis (TBNA) results accurately identified the interactional tasks that were most pertinent to the target speaker group. The practice of conducting a TBNA embodied the principle advocated in Long (2016) where the author argued for the design of more targeted and meaningful teaching and assessment materials that match the needs of the specific end-users groups. The tasks from the TBNA were carefully validated both qualitatively and quantitatively with four groups of informants, drawing on the insight and expertise of L2-Chinese speakers, L1-Chinese speakers, Chinese language teachers and applied linguists (both pragmaticians and testing specialists).

Fifth, this book represents an attempt to democratize applied linguistics and language assessment, decentring the discipline from the few and powerful to the everyday-life people who use language, especially the ones that are historically marginalized. The employment of everyday-life domain experts (DEs) in developing the test construct echoed the advocacy in Hymes (1972) and Sato and McNamara (2019), which highlighted the expertise of everyday-life members in society who use the language, live the language and form opinions and judgements of people who use the language. The IC rating scale and five rating categories represent how language is actually evaluated in real life by real-life people, strengthening the extrapolation inference and making the test more informative to end-users. The CMC platform adopted in this book made test administration much more affordable and allowed me to recruit test-takers from a wide range of walks of life and home countries. More than half of the 26 test-taker countries are recipient countries on the DAC list of ODA (OECD, 2021). This shows that CMC assessment has made IC assessment more accessible to financially disadvantaged L2 speakers who might otherwise not be able to partake in IC language tests and IC assessment research, had the tests been delivered in F2F settings only. The recruitment of test-takers from various languages, ages and

professional backgrounds has also made the research findings more informative and fairer in the sense that the research is not built on knowledge gleaned from affluent middle-class university students from developed countries alone, which features greatly in applied linguistics research.

The sixth contribution is the use of DEs' indigenous criteria to build the IC rating scale and the subsequent theorization of the scale via Sequential-Categorial Analysis, which combines CA and membership categorization analysis (MCA). Different from previous research that extracts differing IC levels based on test-taker performances grouped by proficiency/LC, this book avoided the conflation of LC and IC by resorting to DEs' etic standards, which were uninfluenced by any LC measures. This circumvented the chicken-and-egg question of whether we are truly assessing LC when we claim to be assessing IC. The process of theorizing DEs' indigenous scale into a theorized scale provided the theoretical structure that indigenous criteria frequently lack.

The process of theorization has also revealed the theoretical vitality and robustness of DEs' indigenous criteria as they mirror theories of interpersonal communication in related disciplines such as sociolinguistics and philosophy. Drawing on CA and MCA concepts, the final five rating categories moved IC assessment beyond a mechanistic focus on sequence to social roles, affect, morality and reasoning. Although these dimensions of interaction are under-researched in previous IC assessment studies, they are valued by everyday members in everyday interaction and are emphasized in models of interaction in neighbouring disciplines. The sequential aspect of interaction, which has been the focus of IC assessment so far, has been reconceptualized to be an enabler of the above-mentioned more socially oriented IC dimensions. The assessment of sequential management is therefore incorporated into the assessment of *disaffiliation control*, *affiliation promotion*, *morality*, *reasoning* and *social role management* (see Section 5.2.6).

The cross-disciplinary enrichment in this book echoes the argument in Harding (2021) where he advocated that language testers should reach out to theories and concepts in related fields to expand the scope of the language assessment practice. The highly theorized IC rating scale with its five rating categories suggests stronger applicability across different IC assessment tasks, contexts, and languages. Existing IC rating scales tend to fit closely to their local task contexts, which offers a more precise measurement of IC abilities but lacks generalizability; this makes the development of IC assessment instruments a costly endeavour. The rating scale from this book has a test construct that is broad enough to cover various contexts but can also be localized to assess the specific features of a particular language and the unique interactional features of

particular test tasks. This approach to rating scale design has generated a more comprehensive definition of IC as a test construct and improved the practicality of IC assessment. It has also broadened the applicability of the findings from this book as the theoretical model generated in this book is not specific to the particular target language, particular test types and particular cohorts of test-takers in this research project.

The seventh contribution worth noting is the insight the book offers into the symbiotic relationship between IC and linguistic competence (LC). Using HSK as an LC measure, this book has shown that although there is a connection between IC and LC, this connection is not static. When test-takers are beginner L2 speakers, their limited control over linguistic devices means they are unlikely to attain highly sophisticated interactional performances. At the same time, they are less likely to commit serious IC blunders because of their beginner learner status and the fact that they do not process enough language to commit irrevocable pragmatic mishaps. Beginner L2 speakers, therefore, form a fairly homogenous group in terms of their IC. As their LC improves, test-takers start to demonstrate more agency in their language use and the range of their IC scores increases. Some proficient test-takers can be strong in IC, utilizing their LC to demonstrate competence in handling interpersonal interaction. Some other proficient test-takers can either intentionally or unintentionally commit IC blunders and their strong LC may actually make their blunders more alarming and noticeable. This is most obvious in the L1 group in this book which has the widest range of IC abilities. Although L1 speakers are putatively stronger and more homogenous in LC compared with L2-Chinese speakers, we see some L1-Chinese test-takers perform so much worse than their L2 peers, and sometimes even lag behind lower-proficiency L2 speakers. This finding regarding L1 speakers shows that L1 speakers cannot be safely assumed to be the gold standard when the focus of assessment is IC instead of LC. Not every L1 speaker is strong in IC, and L2 speakers can outperform L1 speakers despite limitations in LC. This finding challenges the privileged position that L1 speakers have enjoyed in language teaching and assessment, highlighting the need to re-evaluate the focus of language teaching and assessment. The nonlinear relationship between IC and LC suggests that IC needs to be taught and assessed in its own right as it is erroneous to assume that stronger LC invariably implies stronger IC. If IC is not emphasized in language classrooms and language tests, we might inadvertently educate language users to become linguistically skilled but interactionally limited. The term 'language users' is used here intentionally because based on the findings from this book it can be argued that the learning and testing of IC is not a task that only L2 speakers

need to undertake. L1 speakers, as shown in this book, can find IC-focused tasks equally challenging and can benefit from the teaching and testing of IC just like L2 speakers do.

The final contribution worth highlighting is the focus this book has on strengthening the extrapolation inference. Weak extrapolation of test results to activities in the TLU domain threatens the validity of the language test and can damage the credibility of the whole language testing industry (O'Sullivan, 2019). Too frequently we are hearing from end-users that test scores are not a good indicator of test-takers' true ability in the TLU domain/*in the wild*, calling into question existing practices in language assessment. To address this concern, this research project put a stronger emphasis on gathering evidence to support the extrapolation inference. Two questionnaires, the self-assessment questionnaire and the peer-assessment questionnaire, were developed and administered, inviting both test-takers and test-taker peers to assess test-takers' IC on each of the five IC rating categories. There was also an additional section in the test-taker self-assessment questionnaire that explicitly asked test-takers to evaluate how valid they thought the test scores reflected their real-world IC conduct. Findings from this extrapolation study have provided insight into how well the test can predicate activities in the TLU domain. The items in the questionnaire also helped us understand test-takers' IC conduct *in the wild* from both the test-taker and the peer perspective, which can inform the teaching and testing of IC.

8.2 Outstanding issues, limitations, and future research

Having summarized the most notable contributions this book makes to the assessment of L2 IC in Section 8.1, Section 8.2 details outstanding issues from this research project, limitations of this project, and directions for future research.

First, the use of only L1-Chinese speakers as DEs to inform the test construct needs to be problematized and discussed. This can be perceived to be problematic given the lingua franca reality of language use: the employment of L1 speakers as the only source of informants for the test construct might be construed as privileging L1 interactional standards and underrepresenting the perspective of L2-Chinese speakers. However, I argue that this practice is necessitated by the aim and context of this project. So far there have not been any studies on test constructs on Chinese IC for general-purpose speaking assessment. For this project, therefore, I decided to first interrogate the IC construct based on L1-speakers' perspective to understand how Chinese IC is understood by L1 speakers, before extending the construct to include the perspective of L2 speakers. A similar approach was adopted by me to look at listening assessment, where

I problematized the convention of using L1-English input in speaking tests and investigated the feasibility of introducing L2-English input to better represent the lingua franca nature of language use (Dai & Roever, 2019b). However, in order to interrogate the validity concerns of assessing L2-English accents, the initial step is to look at the psychometric measures of L1-English input. Therefore, the work in this book represents the first step in unpacking the Chinese IC test construct. Findings of this endeavour show that L1-Chinese speakers converge in their judgement of IC, crystalizing in the five rating categories in the rating scale. Future research can further the work conducted in this book by incorporating the perspective of L2 speakers and revalidating the test construct.

The theory-generating nature of this research project also calls for the use of L1 speakers. When the unifying IC model is still at its nascent stage, it is ideal to recruit construct informants who are strong in both LC and IC. It is unarguable that when other parameters are held constant, L1 speakers in general have stronger LC than L2 speakers due to L1 speakers' longer residence in the target culture and longer period of time in using the target language. The strict selection criteria for L1-Chinese DEs in Section 5.1.1.3 ensures that the DEs selected were strong in both LC and IC, which can better inform the unified IC model. As the resultant IC construct shows, LC plays a role but not a deterministic role in a speaker's IC. Future research can invite L2 speakers to test the vitality of the unified IC model and edit, revise, enrich and expand the model if necessary.

The second point to be problematized about this project is that it stops at the extrapolation inference in the argument-based framework and does not interrogate the decision and consequence inferences (Chapelle, 2020; Knoch & Chapelle, 2017). Both inferences can only be investigated once the test is fully operational to assess whether the decisions made based on the test are fair and whether the test consequences are beneficial to end-users. The use of DEs in this project has, however, paved the way for such investigation as DEs provided insight into the connections between test performance, real-world consequences, and L2-speakers' pedagogical needs in Table 29. Since the test has been fully validated up to the extrapolation inference, future research can build on the findings from this book and further explore the decision and consequence inferences, drawing on the insight from the DEs in Table 29. More research is also needed to connect decision and consequence inferences with the development of stakeholders' assessment literacy, as discussed in Section 7.6.2. The discussion in Section 6.2.5.2 on the test-taker attitude questionnaire items has already shown that the test-takers in this project were viewing IC tests favourably and were dissatisfied with existing speaking tests that prioritize LC. A potentially fruitful line of research moving forward is to examine how test

users' perception of IC assessment changes as their assessment literacy develops, which can generate backings for assumptions for the decision and consequence inferences. Another related topic for future research is the washback of the IC model proposed in this book on pedagogy. The current expanded IC construct suggests that IC is a complex ability that pertains to the sequential, emotional, logical, moral, and categorial dimensions of interpersonal communication. These dimensions are regretfully underemphasized in current language teaching practices. Since assessment is one driving force of what takes place in language classrooms, future research on decision and consequence inferences can look at how language teachers and students take up the IC construct in this book and incorporate it in their curriculum so as to prepare L2 speakers to be not only linguistically competent but also interactionally ready. A discussion on how IC can be taught in classrooms is beyond the scope of this book but interested readers can consult Roever (2022) and Wong and Waring (2020). Readers can also refer to Tai and Dai (2023) to see how the IC model developed in this book was applied to investigate an L1-English ESOL teacher's IC in translanguaging language classroom practices. Dai (2024) is another study that uses the IC model in this book to provide a sociolinguistic-interactional rethink of intercultural communication. The IC model proposed in this book can therefore be used to understand and evaluate all speakers' ability to interact, whether they are language learners or teachers, L1 or L2 speakers, inside or outside the classroom.

Lastly, the use of raters for IC assessment deserves more attention and investigation. This project chose two L1 raters who had no training in applied linguistics, language assessment or IC (Section 6.1.1.3). Both raters went through a carefully designed, scaffolded, three-stage rater training process (Section 6.1.2.4 and 6.1.3.2) and achieved stable ratings in terms of both intra-rater reliability and inter-rater reliability (Section 6.2.1.3). What are the implications here for future rater selection for IC assessment? The answer to this question is twofold: raters' LC and their IC.

For any form of speaking assessment, there is a precondition that the raters for that speaking assessment need to have a high level of LC of the target language to ensure they are capable of comprehending talk from test-takers at different proficiency levels. Even for IC tests that do not directly assess LC, linguistic devices are still the vehicles for the realization of social actions. An inability to appreciate the subtle nuances and differences in word choice or prosodic features can threaten rater reliability and fairness for score users. On this account, both L1 and L2 speakers can serve as IC test raters, as long as the L2 speakers have sufficient LC to cover the range of language the IC test elicits.

The criterion of IC in raters is more complex. Here I hypothesize a difference between raters' meta-IC knowledge and their operational IC. As the findings from this book have shown, L1 speakers are not the gold standard of IC, as many L2 speakers outperformed L1 speakers, though it should also not be overlooked that the highest scores on the test were from L1 speakers (Section 6.2.2). This, however, is test-takers operational IC, which means their ability to implement interactional conducts and is what the IC test was designed for. The IC test cannot assess a person's implicit knowledge of interaction – their knowledge of what is appropriate and what is not. This is termed meta-IC knowledge. My hypothesis is that in terms of IC rater selection, the raters need to have strong meta-IC knowledge but not necessarily strong operational IC. In other words, IC raters need to understand the import and implications of a wide range of interactional patterns and methods, but they do not necessarily need to have the ability or the desire to implement high-IC patterns and methods in real life for themselves. An analogy might make this point clearer. If we need to select good judges for gymnastics, we would like the judges to be able to tell good from bad when examining the performances of a group of gymnasts. It is not mandatory for the judges to have the capacity to achieve what the top-scoring gymnast can do in actual performance. In a similar vein, we would like the IC raters to have strong implicit knowledge of IC but not necessarily the competence to demonstrate strong IC behaviours. Another possible scenario is that the IC raters possess both strong meta-IC knowledge and operational IC, but they may choose to not use their operational IC in real life. As discussed in Section 6.2.2, a person's interactional conducts embody their beliefs, attitudes, values, and personalities. It is very likely some interactants know this is the more polite way to say it (meta-IC knowledge), know how to say it in the more polite way (operational IC), but purposely choose to not say it in the more polite way (actual IC conduct in real life), due to the decisions they make as a free agent. Following this argument, a qualified IC rater does not necessarily have to be able to score high or want to score high in an IC test. As long as they can demonstrate strong meta-IC knowledge and tell a strong IC performance from a not-so-strong one, they should be able to serve as IC raters. If meta-IC knowledge is prioritized over operational IC when it comes to IC rater selection criteria, raters with a higher degree of socialization in the target language community, more extensive experience in the TLU domain, and stronger interactional awareness and aptitude might have an advantage over those who do not. The relationship between these factors and raters' L1/L2 speaker status is unclear and needs to be empirically investigated. L1 speakers may have an advantage over L2 speakers simply because of most L1 speakers' higher degree of socialization in the target

language community and longer history of deployment of the target language. This, however, is largely speculative at this stage and needs to be further researched.

When we re-examine the two raters selected in Chapter 6, both of them were L1 speakers and had extensive experience living, studying and working in the target language. They should have strong LC and meta-IC knowledge, but their operational IC was not measured. Their rating on IC was overall statistically satisfactory. Future research can explore the interconnected relationships between LC, meta-IC knowledge, and operational IC to ascertain the appropriate selection criteria for IC raters.

Another issue related to IC raters is that this project only recruited two raters for the rating of the main testing study. Although both raters demonstrated high rater reliability, future research needs to employ more raters to see if they can still reliably rate IC test performances using the IC scale developed in this book. More research on the cognitive process of IC rating will also be beneficial to better our understanding of raters' meta-IC knowledge and interpretation of the five IC rating categories during rating.

References

Abe, M. (2018, July). *Opening & closing L2 task-based text-chat interactions* [paper presentation]. 5th International Conference on Conversation Analysis, Loughborough, UK.

Abe, M., & Roever, C. (2019). Interactional competence in L2 text-chat interactions: First-idea proffering in task openings. *Journal of Pragmatics, 144*, 1–14. https://doi.org/10.1016/j.pragma.2019.03.001

Abe, M., & Roever, C. (2020). Task closings in L2 text-chat interactions: A study of L2 interactional competence. *CALICO Journal, 37*(1), 23–45. https://doi.org/10.1558/cj.38562

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics, 33*(1), 42–65. https://doi.org/10.1093/applin/amr031

Al-Gahtani, S., & Roever, C. (2014). Preference structure in L2 Arabic requests. *Intercultural Pragmatics, 11*(4), 619–643. https://doi.org/10.1515/ip-2014-0027

Al-Gahtani, S., & Roever, C. (2015). The development of requests by L2 learners of modern standard Arabic: A longitudinal and cross-sectional study. *Foreign Language Annals, 48*(4), 570–583. https://doi.org/10.1111/flan.12157

Al-Gahtani, S., & Roever, C. (2018). Proficiency and preference organization in second language refusals. *Journal of Pragmatics, 129*, 140–153. https://doi.org/10.1016/j.pragma.2018.01.014

Antaki, C. (Ed.). (2016). Orders of interaction in mediated settings [Special Issue]. *Research on Language and Social Interaction, 49*(4).

Aristotle. (2007). *Aristotle on rhetoric: A theory of civic discourse: Translated with Introduction, Notes and Appendices*. (G. A. Kennedy, Trans). Oxford University Press.

Arminen, I., & Weilenmann, A. (2009). Mobile presence and intimacy – Reshaping social actions in mobile contextual configuration. *Journal of Pragmatics*, *41*(10), 1905–1923. https://doi.org/10.1016/j.pragma.2008.09.016

Austin, J. L. (1962). *How to do things with words*. Clarendon.

Balaman, U., & Sert, O. (2017). Development of L2 interactional resources for online collaborative task accomplishment. *Computer Assisted Language Learning, 30*(7), 601–630. https://doi.org/10.1080/09588221.2017.1334667

Bardovi-Harlig, K., & Salsbury, T. (2004). The organization of turns in the disagreements of L2 learners: A longitudinal perspective. In D. Boxer & A. D. Cohen (Eds.), *Studying speaking to inform second language learner* (pp. 199–227). Multilingual Matters.

Bella, S. (2014). Developing the ability to refuse: A cross-sectional study of Greek FL refusals. *Journal of Pragmatics, 61*, 35–62. https://doi.org/10.1016/j.pragma.2013.11.015

Bergmann, J. R. (1998). Introduction: Morality in discourse. *Research on Language & Social Interaction*, *31*(3–4), 279–294. https://doi.org/10.1080/08351813.1998.9683594

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Bloomfield, L. (1931). *Language*. Henry Holt.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Bozkurt, A., Jung, I., Xiao, J., Vladimirschi, V., Schuwer, R., Egorov, G., Lambert, S. R., Al-Freih, M., Pete, J., Olcott Jr., D., Rodes, V., Aranciaga, I., Bali, M., Alvarez Jr., A. V., Roberts, J., Pazurek, A., Raffaghelli, J. E., Panagiotou, N., de Coëtlogon, P., … Paskevicius, M. (2020). A global outlook to the interruption of education due to COVID-19 Pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education, 15*(1), 1–126. https://doi.org/10.5281/zenodo.3878572

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between SLA and language testing research* (pp. 112–114). Cambridge University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25. https://doi.org/10.1191/0265532203lt242oa

Brown, J. D. (2001). *Using surveys in language programmes*. Cambridge University Press.

Brown, J. D. (2009). Foreign and second language needs analysis. In M. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 269–293). Wiley-Blackwell.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language*. Cambridge University Press.

Buttny, R. (1993). *Social accountability in communication*. Sage.

Cabinda, M. (2013). The need for a needs analysis at UEM: Aspects of and attitudes towards change. *Linguistics and Education*, *24*(4), 415–427. https://doi.org/10.1016/j.linged.2013.10.001

CambridgeESOL. (2012). *Cambridge English Business Certificates: Handbook for teachers*. University of Cambridge.

Camras, L. A., Bakeman, R., Chen, Y., Norris, K., & Cain, T. R. (2006). Culture, ethnicity, and children's facial expressions: A study of European American, mainland Chinese, Chinese American, and adopted Chinese girls. *Emotion*, *6*(1), 103–114. https://doi.org/10.1037/1528-3542.6.1.103

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–48. https://doi.org/10.1093/applin/I.1.1

Carnap, R. (1955a). Meaning and synonymy in natural languages. *Philosophical Studies*, *6*(3), 33–47. https://doi.org/10.1007/BF02330951

Carnap, R. (1955b). On some concepts of pragmatics. *Philosophical Studies*, *6*(6), 89–91. https://doi.org/10.1007/BF02341065

CCIS. (2019). 孔子学院研究年度报告2019 [2019 Confucius Institute annual report]. http://www.ccis.sdu.edu.cn/kzxyyjndbg/kzxyyjndbg2019.htm

Cekaite, A. (2007). A child's development of interactional competence in a Swedish L2 classroom. *Modern Language Journal, 91*(1), 45–62. https://doi.org/10.1111/j.1540-4781.2007.00509.x

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), *Intercultural Language Use and Language Learning* (pp. 41–57). Springer.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2), 5–35. https://doi.org/10.5070/L462005216

Chan, A. H., & Lee, P. S. (2005). Intelligibility and preferred rate of Chinese speaking. *International Journal of Industrial Ergonomics, 35*(3), 217–228. https://doi.org/10.1016/j.ergon.2004.09.001

Chan, M. K. M. (1996). Gender-marked speech in Cantonese: The case of sentence-final particles je and jek. *Studies in the Linguistic Sciences, 26*, 1–38.

Chan, M. K. M. (1998). Gender differences in the Chinese language: A preliminary report. In H. Lin (Ed.), *Proceedings of the Ninth North American Conference on Chinese Linguistics* (pp. 35–52). GSIL Publications.

Chapelle, C. (Ed.). (2013). *The encyclopedia of applied linguistics*. Wiley-Blackwell.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. E. Enright & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–350). Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. SAGE Publications.

Chaudron, C., Doughty, C. J., Kim, Y., Kong, D., Lee, J., Lee. Y., Long, M. H., Rivers, R., & Urano, K. (2005). A task-based needs analysis of a tertiary Korean as a foreign language programme. In M. H. Long (Ed.), *Second language needs analysis* (pp. 225–261). Cambridge University Press.

Chen, Y., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDCT and E-mail tasks. *Language Assessment Quarterly*, *13*(3), 231–250. https://doi.org/10.1080/15434303.2016.1213844

Chinesetest. (n.d.). *Chinese Proficiency Test (HSK)*. http://www.chinesetest.cn/gosign.do?id=1&lid=0

Chu, C. C. (1990). Semantics and discourse in Chinese language instruction. *Journal of the Chinese Language Teachers' Association, 25*, 15–29.

Clayman, S. E. (2002). Sequence and solidarity. In E. J. Lawler & S. R. Thye (Eds.), *Advances in group processes: Group cohesion, trust, and solidarity* (pp. 229–253). Elsevier Science.

Coleman, J. S. (1990). *Foundations of social theory*. Harvard University Press.

Confucius Institute/Hanban. (2017). *Confucius Institute annual development report 2017*. Hanban. http://www.hanban.edu.cn/report/2017.pdf

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume with new descriptors.* Council of Europe.

Couper-Kuhlen, E. (2012). Exploring affiliation in the reception of conversational complaint stories. In A. Peräkylä & M. L. Sorjonen (Eds.), *Emotion in interaction* (pp. 113–146). Oxford University Press.

Cowling, J. D. (2007). Needs analysis: Planning a syllabus for a series of intensive workplace courses at a leading Japanese company. *English for Specific Purposes*, *26*(4), 426–442. https://doi.org/10.1016/j.esp.2006.10.003

Crystal, D. (2011). *Internet linguistics*. Routledge.

Curl, T. S., & Drew, P. (2008). Contingency and action: A comparison of two forms of requesting. *Research on Language and Social Interaction, 41*(2), 129–153. https://doi.org/10.1080/08351810802028613

Dai, D. W. (2019a, June 9–14). "*Wang Si sister hello!" - social role enactment and disaffiliation management by L2 and L1 Chinese speakers*. [Conference session]. 16th International Pragmatics Conference (IPrA), Hong Kong, China. https://cdn.ymaws.com/pragmatics.international/resource/collection/C57D1855-A3BB-40D8-A977-4732784F7B21/IPRA2019_Abstracts_Book-with_corrections.pdf

Dai, D. W. (2019b, November 25–27). *Where non-native speakers outperformed native speakers: Interrogating the indigenous criteria for appropriate language use in a Chinese speaking test*. [Conference session]. The 2019 Applied Linguistics Association of Australia (ALAA), Applied Linguistics in Aotearoa New Zealand (ALANZ) and Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) Conference, Perth, WA, Australia. http://researcharchive.wintec.ac.nz/7220/1/ALAA-ALAN%20program%20online%20-%20191119.pdf

Dai, D. W. (2019c, November 25–27). "The way he talks is really obnoxious": Hearing what listener judges cannot hear with Conversation Analysis and Membership Categorisation Analysis. [Conference session]. The 2019 Applied Linguistics Association of Australia (ALAA), Applied Linguistics in Aotearoa New Zealand (ALANZ) and Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) Conference, Perth, WA, Australia. http://researcharchive.wintec.ac.nz/7220/1/ALAA-ALAN%20program%20online%20-%20191119.pdf

Dai, D. W. (2021, March 20–23). *Design and validation of a speaking rubric measuring L2 Chinese interactional competence*. [Conference session]. American Association for Applied Linguistics (AAAL) 2021 Conference, virtual conference. https://www.xcdsystem.com/aaal/program/64O29Sh/index.cfm

Dai, D. W. (2022). *Design and validation of an L2-Chinese interactional competence test*. [Doctoral dissertation, University of Melbourne, Australia]. https://www.proquest.com/openview/fda5f389be0274fe555b4ff6275231dc/1?pq-origsite=gscholar&cbl=18750&diss=y

Dai, D. W. (2023a). What do second language speakers really need for real-world interaction? A needs analysis of L2 Chinese interactional competence. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688221144836

Dai, D. W. (2023b, June 13–15). *Using Sequential-Categorial Analysis to assess interactional competence*. [Pre-conference workshop]. The 19th European Association for Language Testing and Assessment (EALTA) Conference, Helsinki, Finland. https://www.helsinki.fi/en/conferences/ealta-conference-2023/pre-conference-workshops

Dai, D. W. (2023c). *Defining and assessing interactional competence*. [Conference session]. American Association for Applied Linguistics (AAAL) 2023 Conference, Portland, OR, United States. https://www.researchgate.net/publication/372338843_Defining_and_assessing_Interactional_Competence

Dai, D. W. (2024). Interactional Competence for professional communication in intercultural contexts: Epistemology, analytic framework and pedagogy. *Language, Culture and Curriculum*. https://doi.org/10.1080/07908318.2024.2349781

Dai, D. W., Grieve, A., & Yahalom, S. (2022a). Interactional competence in the online space: Affordances, challenges, and opportunities for TESOL practitioners. [Editorial]. *TESOL in Context, 30*(2), 1–7. https://doi.org/10.21153/tesol2022vol30no2art1703

Dai, D. W., Grieve, A., & Yahalom, S. (Eds.). (2022b). Interactional competence in the online space. [Special issue]. *TESOL in Context, 30*(2).

Dai, D. W., & Davey, M. (2022, September 8–10). *Doing being a social member: Membership categorisation practices as an indicator of interactional competence*. [Conference session]. The 2022 Interactional Competences and Practices in a Second Language (ICOP-L2) Conference, Barcelona, Spain. https://e7c0b21e-6f8a-41e6-8cff-98a52b01cef9.filesusr.com/ugd/d3edd2_9f753ee5197545f5a96f4836402819dd.pdf

Dai, D. W., & Davey, M. (2023). On the promise of using Membership Categorization Analysis to investigate Interactional Competence. *Applied Linguistics*. https://doi.org/10.1093/applin/amad049

Dai, D. W., & Roever, C. (2019a, March 9–12). *A task-based needs analysis of interactional competence in Chinese as a foreign language*. [Conference session]. American Association for Applied Linguistics (AAAL) 2019 Conference, Atlanta, GA, USA. https://aaal.confex.com/aaal/2019/meetingapp.cgi/Session/1616

Dai, D. W., & Roever, C. (2019b). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Language Assessment Quarterly*, *16*(1), 64–86. https://doi.org/10.1080/15434303.2019.1601198

Davidson, D. (1980). *Essays on actions and events*. Clarendon Press.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177%2F0265532215582282

Diao, W. (2014). Peer socialization into gendered L2 Mandarin practices in a study abroad context: Talk in the dorm. *Applied Linguistics*, *37*(5), 599–620. https://doi.org/10.1093/applin/amu053

Dings, A. (2014). Interactional competence and the development of alignment activity. *The Modern Language Journal, 98*(3), 742–756. https://doi.org/10.1111/modl.12120

Dinkins, C. S. (2005). Shared inquiry: Socratic-hermeneutic interpre-viewing. In P. M. Ironside (Ed.), *Beyond method: Philosophical conversations in healthcare research and scholarship* (pp. 111–147). The University of Wisconsin Press.

Donagan, A. (1964). Historical explanation: The Popper-Hempel theory reconsidered. *History and Theory*, *4*(1), 3–26. https://doi.org/10.2307/2504200

Dooly, M., & Sadler, R. (2016). Becoming little scientists: Technologically-enhanced project-based language learning. *Language Learning & Technology, 20*(1), 54–78.

Drew, P. (1984). Speakers' reportings in invitation sequences. In J. M. Atkinson & J. C. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 129–151). Cambridge University Press.

Drew, P. (2013). Turn design. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 131–149). Wiley-Blackwell.

Drew, P., Ogden, R., & Curl, T. (2006). *Affiliation and disaffiliation in interaction: Language and social cohesion*. ESRC RES-00023-0035 End-of-award report. https://www.researchcatalogue.esrc.ac.uk/grants/RES-000-23-0035/outputs/read/c537444a-5c46-40f0-b8ba-7e69fe48ea31

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177%2F0265532209104669

Dugartsyrenova, V. A., & Sardegna, V. G. (2017). Developing oral proficiency with VoiceThread: Learners' strategic uses and views. *ReCALL*, *29*(1), 59–79. https://doi.org/10.1017/S0958344016000161

Duranti, A. (1997). *Linguistic anthropology*. Cambridge University Press.

Eggly, S., Musial, J., & Smulowitz, J. (1999). The relationship between English language proficiency and success as a medical resident. *English for Specific Purposes, 18*, 201–208. https://doi.org/10.1016/S0889-4906(98)00002-7

Elder, C., & McNamara, T. (2016). The hunt for 'indigenous criteria' in assessing communication in the physiotherapy workplace. *Language Testing, 33*(2), 153–174. https://doi.org/10.1177%2F0265532215607398

Farris, C. S. (1988). Gender and grammar in Chinese: With implications for language universals. *Modern China*, *14*(3), 277–308. https://doi.org/10.1177%2F009770048801400302

Farris, C. S. (1994). A semiotic analysis of sajiao as a gender marked communication style in Chinese. In M. Johnson & F. Y. L. Chiu (Eds.), *Unbound Taiwan: Close-ups from a distance* (pp. 1–29). University of Chicago.

Felix-Brasdefer, J. C. (2012). E-mail requests to faculty: E-politeness and internal modification. In M. Economidou-Kogetsidis & H. Woodfield (Eds.), *Interlanguage request modification* (pp. 87–118). John Benjamins.

Firth, J. R. (1955). Structural linguistics. *Transactions of the Philological Society, 54*(1), 83–103. https://doi.org/10.1111/j.1467-968X.1955.tb00290.x

Fitzgerald, R., & Housley, W. (Eds.). (2015). *Advances in membership categorization analysis*. Sage.

Fitzgerald, R., Housley, W., & Rintel, S. (Eds.). (2017). Membership Categorization Analysis. Technologies of Social Action [Special Section]. *Journal of Pragmatics, 118*, 51–133.

Føllesdal, D. (1986). Intentionality and rationality. In J. Margolis, M. Krausz & R. M. Burian (Eds.), *Rationality, relativism and the human sciences* (pp. 109–125). Springer.

Francis, D., & Hart, C. (1997). Narrative intelligibility and membership categorization in a television commercial. In S. Hester & P. Eglin (Eds.), *Culture in action: Studies in membership categorization analysis* (pp. 123–52). University Press of America.

Freud, S. (1956). *A general introduction to psychoanalysis*. Permabooks.

Fulcher, G. (2003). *Testing second language speaking*. Longman, Pearson Education.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*, 5–29. https://doi.org/10.1177%2F0265532209359514

Galaczi, E. D. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gardner, R., & Wagner, J. (Eds.). (2004). *Second language conversations*. Continuum.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall.

Gilabert, R. (2005). Evaluating the use of multiple sources and methods in needs analysis: a case study of journalists in the autonomous community of Catalonia (Spain). In M. H. Long (Ed.), *Second language needs analysis* (pp. 182–199). Cambridge University Press.

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry: Journal of Interpersonal Relations, 18*, 213–31. https://doi.org/10.1080/00332747.1955.11023008

Goffman, E. (1956). The nature of deference and demeanor. *American Anthropologist, 58*, 473–502. https://doi.org/10.1525/aa.1956.58.3.02a00070

Goffman, E. (1963). *Behaviour in public places*. Free Press.

Goffman, E. (1964). The neglected situation. *American Anthropologist, 66* (6), Part 2, 133–136. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00090

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior*. Doubleday, Anchor Books.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harper & Row.

Golato, A. (2002). German compliment responses. *Journal of Pragmatics, 34*, 547– 571. https://doi.org/10.1016/S0378-2166(01)00040-6

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, *24*, 90–121. https://doi.org/10.1093/applin/24.1.90

Gonzales, A. (2013). Development of politeness strategies in participatory online environments: a case study. In N. Taguchi & J. M. Sykes (Eds.), *Language learning & language teaching: Vol. 36. Technology in interlanguage pragmatics research and teaching* (pp. 101–120). John Benjamins.

González-Lloret, M. (2011). Conversation analysis of computer-mediated communication. *CALICO Journal*, *28*(2), 308–325. http://dx.doi.org/10.11139/cj.28.2.308-325

González-Lloret, M. (2015). Conversation analysis in computer-assisted language learning. *CALICO Journal*, *32*(3), 569–595. http://dx.doi.org/10.1558/cj.v32i3.27568

Grandy, R. E. (1989). On Grice on language. *The Journal of Philosophy*, *86*(10), 514–525.

Greer, T. (2016). Learner initiative in action: Post-expansion sequences in a novice ESL survey interview task. *Linguistics and Education, 35*, 78–87. https://doi.org/10.1016/j.linged.2016.06.004

Greer, T. (2020). Using a conversation analytic exemplar-based rubric to assess engagement in a paired EFL test. 神戸大学国際コミュニケーションセンター論集, (16), 1–24.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388.

Grice, P. (1975). Language and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41–58). Academic Press.

Gruba, P., Jin, Y., Laborda, J. G., Purpura, J., Davoodifard, M., Green, T., & Chapelle, C. A. (2020, June). *Exploring the future of online language testing and assessment: Technological opportunities and challenges*. ILTA Webinar.

Habermas, J. (1981). *A theory of communicative action*. Beacon Press.

Hall, J. K., & Pekarek Doehler, S. (2011). L2 interactional competence and development. In J. K. Hall, J. Hellermann & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 1–18). Multilingual Matters.

Hall, P., Keely, E., Dojeiji, E., Byszewski, A., & Marks, M. (2004). Communication skills, cultural challenges and individual support: Challenges of international medical graduates in a Canadian healthcare environment. *Medical Teacher, 26*(2), 120–125. https://doi.org/10.1080/01421590310001653982

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Harding, L. [SALT Journal at Teachers College, CU]. (2021, Jan). *Language assessment: Current trends, future challenges* [Youtube Video]. Youtube. https://www.youtube.com/watch?v=UljPM8H9ywI&t=200s

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago University Press.

Haugh, M., & Obana, Y. (2015). Transformative continuations, (dis)affiliation, and accountability in Japanese interaction. *Text & Talk, 35*(5), 597–619. https://doi.org/10.1515/text-2015-0015

Hayashi, M. (2013). Turn allocation and turn sharing. In J. Sidnell & T. Stivers (Eds.), *Handbook of conversation analysis* (pp. 167–190). Wiley-Blackwell.

Hellermann, J. (2007). The development of practices for action in classroom dyadic interaction: Focus on task openings. *The Modern Language Journal, 91*(1), 83–96. https://doi.org/10.1111/j.1540-4781.2007.00503.x

Hellermann, J. (2008). *Social actions for classroom language learning*. Multilingual Matters.

Hellermann, J. (2009). Practices for dispreferred responses using no by a learner of English. *International Review of Applied Linguistics in Language Teaching, 47*, 95–126. https://doi.org/10.1515/iral.2009.005

Hellermann, J., Eskildsen, S. W., Doehler, S. P., & Piirainen-Marsh, A. (Eds.). (2019). *Conversation analytic research on learning-in-action: The complex ecology of second language interaction 'in the wild'* (Vol. 38). Springer Nature.

Hempel, C. G. (1961). Rational action. *Proceedings and Addresses of the American Philosophical Association, 35*, 5–23. https://doi.org/10.2307/3129344

Heritage, J. (1984). *Garfinkel and ethnomethodology*. Polity Press.

Heritage, J. (1988). Explanations as accounts: A conversation analytic perspective. In C. Antaki (Ed.), *Analysing everyday explanation: A casebook of methods* (pp. 127–144). Sage.

Heritage, J. (1990). Intention, meaning and strategy: Observations on constraints on interaction analysis. *Research on Language and Social Interaction, 24*(1–4), 311–332. https://doi.org/10.1080/08351819009389345

Heritage, J. (2011). Conversation analysis: Practices and methods. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 208–230). Sage.

Herring, S. C. (2013). Discourse in Web 2.0: Familiar, reconfigured, and emergent. In D. Tannen & A. M. Trester (Eds.), *Georgetown University Round Table on languages and linguistics 2011: Discourse 2.0: Language and new media,* (pp. 1–25). Georgetown University Press.

Hester, S., & Eglin, P. (Eds.). (1997). *Culture in action: Studies in membership categorization analysis*. University Press of America.

Hofstede Insight. (2021). *Comparing countries.* https://www.hofstede-insights.com/product/compare-countries/

Holme, R., & Chalauisaeng, B. (2006). The learner as needs analyst: The use of participatory appraisal in the EAP reading classroom. *English for Specific Purposes*, *25*(4), 403–419. https://doi.org/10.1016/j.esp.2006.01.003

Homans, G. (1961). *Social behavior: Its elementary forms*. Harcourt Brace Jovanovich.

Huh, S. (2006). A task-based needs analysis for a business English course. *Second Language Studies*, *24*(2), 1–64. https://www.hawaii.edu/sls/wp-content/uploads/2014/09/HuhSorin.pdf

Hung, Y. W., & Higgins, S. (2016). Learners' use of communication strategies in text-based and video-based synchronous computer-mediated communication environments: Opportunities for language learning. *Computer Assisted Language Learning*, *29*(5), 901–924. https://doi.org/10.1080/09588221.2015.1074589

Hutchby, I., & Tanna, V. (2008). Aspects of sequential organization in text message exchange. *Discourse and Communication*, *2*, 143–64. https://doi.org/10.1177%2F1750481307088481

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Huth, T. (2006). Negotiating structure and culture: L2 learners' realization of L2 compliment-response sequences in talk-in-interaction. *Journal of Pragmatics*, *38*(12), 2025–2050. https://doi.org/10.1016/j.pragma.2006.04.010

Huth, T., & Taleghani-Nikazm, C. (2006). How can insights from conversation analysis be directly applied to teaching L2 pragmatics? *Language Teaching Research*, *10*(1), 53–79. https://doi.org/10.1191%2F1362168806lr184oa

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. University of Pennsylvania Press.

Hymes, D. (l967). Models of the interaction of language and social setting. *Journal of Social Issues, 23*(2), 8–38. https://doi.org/10.1111/j.1540-4560.1967.tb00572.x

IELTS. (n.d.). *SPEAKING: Band Descriptors (public version)*. https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

Ikeda, N. (2017). *Measuring L2 oral pragmatic abilities for use in social contexts: Development and validation of an assessment instrument for L2 pragmatics performance in university settings*. [Unpublished Ph.D. dissertation, University of Melbourne].

Ikeda, N. (2021). Assessing L2 learners' pragmatic ability in problem-solving situations at English-medium university. *Applied Pragmatics, 3*(1), 51–83. https://doi.org/10.1075/ap.19039.ike

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600–619. https://doi.org/10.1177%2F0265532220943483

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jacoby, S. W. (1998). Science as performance: Socializing scientific discourse through conference talk rehearsals. [Unpublished doctoral dissertation, University of California, Los Angeles].

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating guides for TOEFL IBT's integrated speaking tasks.* ETS Research Report. https://www.ets.org/Media/Research/pdf/RR-13-13.pdf

Jasso-Aguilar, R. (2005). Sources, methods and triangulation in needs analysis: A critical perspective in a case study of Waikiki hotel maids. In M. H. Long (Ed.), *Second language needs analysis* (pp. 127–158). Cambridge University Press.

Jayyusi, L. (1984). *Categorization and the moral order*. Routledge and Kegan Paul.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). John Benjamins.

Jenks, C. J. (2014). *Social interaction in second language chat rooms.* Edinburgh University Press.

Jenks, C. J., & Brandt, A. (2013). Managing mutual orientation in the absence of physical copresence: Multiparty voice-based chat room interaction. *Discourse Processes, 50*(4), 227–248. https://doi.org/10.1080/0163853X.2013.777561

Johnson, M. (2001). *The art of nonconversation*. Yale University Press.

Junn, H. (2021). L2 communicative competence analysis via synchronous computer-mediated communication (SCMC) as an alternative to formal classrooms. *Innovation in Language Learning and Teaching*, 1–17. https://doi.org/10.1080/17501229.2021.1895802

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Greenwood Publishing.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17. https://doi.org/10.1177%2F0265532211417210

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K. Bardovi-Harlig, C. Félix-Brasdefer & A. S. Omar (Eds.), *Pragmatics and language learning* (Vol. 11, pp. 281–314). Second Language Teaching and Curriculum Center, University of Hawai'i.

Kasper, G., & Wagner, J. (2011). A conversation-analytic approach to second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 117–142). Routledge.

Ke, C. (2012). Research in second language acquisition of Chinese: Where we are, where we are going. *Journal of Chinese Language Teachers' Association, 47*, 43–113.

Kendrick, K. H. (2018). Adjusting epistemic gradients: The final particle ba in Mandarin Chinese conversation. *East Asian Pragmatics*, 5–26. http://dx.doi.org/10.1558/eap.36120

Kim, Y. K. (2013). Frame analysis. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Blackwell.

Kim, H. Y. (2017). Effect of modality and task type on interlanguage variation. *ReCALL*, *29*(2), 219–236. https://doi.org/10.1017/S0958344017000015

Kley, K. (2019). What counts as evidence for interactional competence? Developing rating criteria for a German classroom-based paired speaking test. In S. Kunitz & R. Salaberry (Eds.), *Teaching and testing L2 interactional competence: Bridging theory and practice* (pp. 291–321). Routledge.

Knoch, U., & Macqueen, S. (2020). *Assessing English for professional purposes*. Routledge.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, 1–23. https://doi.org/10.1177%2F0265532217710049

Knoch, U., & Elder, C. (2013). A framework for validating post entry language assessment. *Papers in Language Testing and Assessment*, *2*(2), 48–66. http://www.altaanz.org/uploads/5/9/0/8/5908292/4_knoch_elder.pdf

Knoch, U. (2012). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33*(1), 77–86. https://doi.org/10.1016/j.esp.2013.08.002

Knoch, U., Elder, C., Woodward-Kron, R., Manias, E., Flynn, E., McNamara, T., Huisman, A., & Zhang, B. Y. (2020). Capturing domain expert perspectives in devising a rating scale for a health specific writing test: How close can we get? *Assessing Writing*, *46*, 1–13. https://doi.org/10.1016/j.asw.2020.100489

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, 1–25. https://doi.org/10.1177/0265532221994052

Kohnke, L., & Moorhouse, B. L. (2020). Facilitating synchronous online language learning through Zoom. *RELC Journal*, 1–6. https://doi.org/10.1177%2F0033688220937235

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*, 366–372. https://doi.org/10.1111/j.1540-4781.1986.tb05291.x

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the

language assessment literacy survey. *Language Assessment Quarterly, 17*(1), 100–120. https://doi.org/10.1080/15434303.2019.1674855

Lambert, C. (2010). A task-based needs analysis: Putting principles into practice. *Language Teaching Research*, *14*(1), 99–112. https://doi.org/10.1177%2F1362168809346520

Lan, L., & MacGregor, L. (2010). English in tiers in the workplace: A case study of email usage. In G. Forey & J. Lockwood (Eds.), *Globalization, communication and the workplace: Talking across the world* (pp. 8–24). Continuum.

Laughlin, V. T., Wain, J., & Schmidgall, J. (2015). *Defining and operationalizing the construct of pragmatic competence: Review and recommendations*. ETS research report. https://doi.org/10.1002/ets2.12053

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*(3), 373–386. https://doi.org/10.1016/0346-251X(92)90047-7

Lee, Y. A., & Hellermann, J. (2014). Tracing developmental changes through conversation analysis: Cross-sectional and longitudinal analysis. *TESOL Quarterly, 48*(4), 763–788. https://doi.org/10.1002/tesq.149

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford University Press.

Levinson, S. C. (2006). On the human 'interaction engine'. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality* (pp. 39–69). Berg.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*, 731. https://doi.org/10.3389/fpsyg.2015.00731

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Li, S. (2013). The role of instruction in developing pragmatic competence in L2 Chinese. In Z. Jing-Schmidt (Ed.), *Increased empiricism: Recent advances in Chinese linguistics* (pp. 293–308). John Benjamins Publishing Company.

Licoppe, C., & Morel, J. (2012). Video-in-interaction: 'Talking heads' and the multimodal organization of mobile and Skype video calls. *Research on Language & Social Interaction*, *45*(4), 399–429. https://doi.org/10.1080/08351813.2012.724996

Lim, G. S. (2018). Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century. *Language Assessment Quarterly, 15*(3), 215–218. https://doi.org/10.1080/15434303.2018.1482493

Lim, N. E. (2011). From subjectivity to intersubjectivity: Epistemic marker wo juede in Chinese. In Y. Xiao, L. Tao & H. L. Soh (Eds.), *Studies in Chinese Linguistics in the New Era* (pp. 265–300). Cambridge Scholars Press.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions, 12*, 636.

Linacre, J. M. (1997). Kr-20/Cronbach alpha or rasch person reliability: which tells the 'truth'? *Rasch Measurement Transactions, 11*(3), 580–1.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement, 3*(1), 85–106.

Lindström, A., & Sorjonen, M.-L. (2013). Affiliation in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 350–369). Wiley-Blackwell.

Linnell, J. D. (2001). Chinese as a second/foreign language teaching and research: Changing classroom contexts and teacher choices. *Language Teaching Research*, *5*(1), 54–81. https://doi.org/10.1177%2F136216880100500104

Locher, M. A. (2014). Electronic discourse. In K. P. Schneider & A. Barron (Eds.), *Pragmatics of Discourse* (pp. 555–581). De Gruyter.

Loewen, S., & Isbell, D. R. (2017). Pronunciation in face-to-face and audio-only synchronous computer-mediated learner interactions. *Studies in Second Language Acquisition*, *39*(2), 225–256. https://doi.org/10.1017/S0272263116000449

Long, M. H. (2005a). A rationale for needs analysis and needs analysis research. In M. H. Long (Ed.), *Second language needs analysis* (pp. 1–16). Cambridge University Press.

Long, M. H. (2005b). Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis* (pp. 19–76). Cambridge University Press.

Long, M. H. (2013a). Identifying language needs for TBLT in the tourist industry. In G. Bosch (Ed.), *Teaching foreign languages for tourism: Research and practice* (pp. 21–44). Peter Lang.

Long, M. H. (2013b). Needs analysis. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell.

Long, M. H. (2015a). Task-based syllabus design. In M. H. Long (Ed.), *Second language acquisition and task-based language teaching* (pp. 205–247). Wiley-Blackwell.

Long, M. H. (2015b). Identifying target tasks. In M. H. Long (Ed.), *Second language acquisition and task-based language teaching* (pp. 117–168). Wiley-Blackwell.

Long, M. H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5–33. https://doi.org/10.1017/S0267190515000057

Ma, X., Gong, Y., Gao, X., & Xiang, Y. (2017). The teaching of Chinese as a second or foreign language: A systematic review of the literature 2005–2015. *Journal of Multilingual and Multicultural Development*, *38*(9), 815–830. https://doi.org/10.1080/01434632.2016.1268146

Malicka, A., Guerrero, R. G., & Norris, J. M. (2017). From needs analysis to task design: Insights from an English for specific purposes context. *Language Teaching Research*, 1–29. https://doi.org/10.1177%2F1362168817714278

Malone, M. E., & Montee, M. (2014). *Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability*. ETS Research Report. https://doi.org/10.1002/ets2.12039

Martin, A., & Adrada-Rafael, S. (2017). Business Spanish in the real world: A task-based needs analysis. *L2 Journal*, *9*(1), 39–61. https://doi.org/10.5070/L29131409

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–421. https://doi.org/10.1177%2F0265532209104668

May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing*, *37*(2), 165–188. https://doi.org/10.1177%2F0265532219879044

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–242. https://doi.org/10.1017/S0267190502000120

McNamara, T. (2014). 30 years on – Evolution or revolution? *Language Assessment Quarterly, 11*(2), 226–232. https://doi.org/10.1080/15434303.2014.895830

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment*. Oxford University Press.

Mehan, H. (1979). *Learning lessons.* Harvard University Press.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, *35*, 41–55. https://doi.org/10.1016/j.asw.2017.12.003

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mochizuki, N. (2017). Contingent needs analysis for task implementation: An activity systems analysis of group writing conferences. *TESOL Quarterly*, *51*(3), 607–631. https://doi.org/10.1002/tesq.391

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia, 10*, 1–16. https://doi.org/10.1186/s40468-020-00101-6

Molle, D., & Prior, P. (2008). Multimodal genre systems in EAP writing pedagogy: Reflecting on a needs analysis. *TESOL Quarterly, 42*(4), 541–566. https://doi.org/10.1002/j.1545-7249.2008.tb00148.x

Moorhouse, B. L., & Beaumont, A. M. (2020). Utilizing video conferencing software to teach young language learners in Hong Kong during the COVID-19 class suspensions. *TESOL Journal, 11*(3). https://doi.org/10.1002/tesj.545

Moorhouse, B. L., Li, Y., & Walsh, S. (2021). E-Classroom interactional competencies: Mediating and assisting language learning during synchronous online lessons. *RELC Journal*, 1–15. https://doi.org/10.1177%2F00336882209852 74

Morris, C. (1938). Foundations of the theory of signs. In O. Neurath, R. Carnap & C. Morris (Eds.), *International encyclopedia of unified science* (pp. 77–138). University of Chicago Press.

Murray, J. C., Cross, J. L., & Cruickshank, K. (2014). *Stakeholder perceptions of IELTS as a gateway to the professional workplace: The case of employers of overseas trained teachers*. IELTS Research Reports Online Series. https://www.ielts.org/for-researchers/research-reports/online-series-2014-1

Negretti, R. (1999). Web-based activities and SLA: A conversational analysis approach. *Language Learning & Technology, 3*(1), 75–87. https://doi.org/10125/25057

Nguyen, M. H., Gruber, J., Fuchs, J., Marler, W., Hunsaker, A., & Hargittai, E. (2020). Changes in digital communication during the COVID-19 global pandemic: Implications for digital inequality and future research. *Social Media + Society,* 6(3), 1–6. https://doi.org/10.1177%2F2056305120948255

Nguyen, T. T. M. (2018). Pragmatic development in the instructed context. *Pragmatics, 28*(2), 217–252. https://doi.org/10.1075/prag.00007.ngu

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15(2)*, 217–263. https://doi.org/10.1177%2F026553229801500204

O'Sullivan, B. (2019). The future of language testing. In C. Roever & G. Wigglesworth (Eds.), *Social perspectives on language testing* (pp. 197–216). Peter Lang.

Ochs, E. (1984). Clarification and culture. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 325–341). Georgetown University Press.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39–62. https://doi.org/10.1177%2F0265532214538014

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly, 18*(1), 1–5. https://doi.org/10.1080/15434303.2020.1866576

OECD. (2021). *DAC list of ODA recipients: Effective for reporting on 2021 flows*. https://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/DAC-List-ODA-Recipients-for-reporting-2021-flows.pdf

Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly, 17*(1), 60–84. https://doi.org/10.1080/15434303.2019.1674854

Oliver, R., Grote, E., Rochecouste, J., & Exell, M. (2013). Needs analysis for task-based language teaching: A case study of indigenous vocational education and training students who speak EAL/EAD. *TESOL in Context*, *22*(2), 36–50.

Oller, J. (1979). *Language tests at school*. Longman.

Orton, J. (2014). Comparing teachers' judgments of learners' speech in Chinese as a foreign language. *Foreign Language Annals*, *47*(3), 507–526. https://doi.org/10.1111/flan.12097

Park, M., & Slater, T. (2014). A typology of tasks for mobile-assisted language learning: Recommendations from a small-scale needs analysis. *TESL Canada Journal, 31*(8), 93–115. https://doi.org/10.18806/tesl.v31i0.1188

Parsons, T. (1937). *The structure of social action*. McGrawHill.

Payne, G. C. F. (1976). Making a lesson happen: An ethnomethodological analysis. In M. Hammersley & P. Woods (Eds.), *The process of schooling*. Open University Press.

Pearson Education. (2020, April). *PTE Academic, Score Guide for institutions, Version 4*. https://pearsonpte.com/wp-content/uploads/2020/04/Score-Guide-16.04.19-for-institutions.pdf

Pekarek Doehler, S., & Berger, E. (2016). L2 interactional competence as increased ability for context-sensitive conduct: A longitudinal study of story-openings. *Applied Linguistics*, 1–25. https://doi.org/10.1093/applin/amw021

Pekarek Doehler, S., & Pochon-Berger, E. (2011). Developing 'methods' for interaction: A cross-sectional study of disagreement sequences in French

L2. In J. K. Hall, J. Hellermann & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 206–243). Multilingual Matters.

Pekarek Doehler, S., & Pochon-Berger, E. (2015). The development of L2 interactional competence: Evidence from turn-taking organization, sequence organization, repair organization and preference organization. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 233–268). Mouton de Gruyter.

Pekarek Doehler, S., Wagner, J., & González-Martínez, E. (2018). *Longitudinal studies on the organization of social interaction*. Palgrave macmillan.

Pekarek Doehler, S. (2018). Elaborations on L2 interactional competence: The development of L2 grammar-for-interaction. *Classroom Discourse, 9*(1), 3–24. https://doi.org/10.1080/19463014.2018.1437759

Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing, 38*(2), 326–337. https://doi.org/10.1177%2F0265532220957298

Piirainen-Marsh, A., & Tainio, L. (2014). Asymmetries of knowledge and epistemic change in social gaming interaction. *The Modern Language Journal, 98*(4). 1022–1038. https://doi.org/10.1111/modl.12153

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33*(2), 175–193. https://doi.org/10.1177%2F0265532215607400

Plough, I., Banerjee, J., & Iwashita, N. (Eds.). (2018). Special issue on interactional competence [Special issue]. *Language Testing*, *35*(3).

Pojanapunya, P., & Jaroenkitboworn, K. (2011). How to say 'Good-bye' in Second Life. *Journal of Pragmatics*, *43*(14), 3591–3602. https://doi.org/10.1016/j.pragma.2011.08.010

Polat, B. (2011). Investigating acquisition of discourse markers through a developmental learner corpus. *Journal of Pragmatics, 43*(15), 3745–3756. https://doi.org/10.1016/j.pragma.2011.09.009

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Studies in Language Testing*, *3*, 74–91.

Pomerantz, A., & Heritage, J. (2013). Preference. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 210–228). Wiley-Blackwell.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. C. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–111). Cambridge University Press.

Pomerantz, A. (1986). Extreme case formulations: A way of legitimizing claims. *Human Studies*, *9*, 219–229. https://doi.org/10.1007/BF00148128

Popper, K. (1957). *The open society and its enemies*. Routledge.

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*(1), 1012.

Robinson, J. D. (2013). Overall structural organization. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 257–280). Wiley Blackwell.

Roever, C., & Dai, D. W. (2021). Reconceptualising interactional competence for language testing. In M. R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context: Expanding the construct and its applications* (pp. 23–49). Multilingual Matters. https://doi.org/10.21832/9781788923828-003

Roever, C., & Ikeda, N. (2021). What scores from monologic speaking tests can (not) tell us about interactional competence. *Language Testing*, 1–23. https://doi.org/10.1177%2F02655322211003332

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331–355. https://doi.org/10.1177%2F0265532218758128

Roever, C. (2005). *Testing ESL pragmatics*. Peter Lang.

Roever, C., Fraser, C., & Elder, C. (2014). *Testing ESL sociopragmatics: Development and validation of a web-based test battery*. Peter Lang.

Roever, C. (2018, May). *Assessing interactional competence. Features, scoring, and practicality* [Keynote conference address]. Assessing Speaking in Context – New Trends, Rice University, Houston, TX, USA.

Roever, C. (2022). *Teaching and testing second language pragmatics and interaction: A practical guide*. Routledge.

Ross, S. (2018). Listener response as a facet of interactional competence. *Language Testing*, *35*(3), 357–375. https://doi.org/10.1177%2F0265532218758125

Rouhshad, A., Wigglesworth, G., & Storch, N. (2016). The nature of negotiations in face-to-face versus computer-mediated communication in pair interactions. *Language Teaching Research*, *20*(4), 514–534. https://doi.org/10.1177%2F1362168815584455

Roulston, K. (2010). *Reflective interviewing: A guide to theory and practice*. Sage.

Ruusuvuori, J. (2013). Emotion, affect and conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 330–349). Wiley-Blackwell.

Rylander, J. (2011). An introduction to needs analysis methods: Options and feasibility issues. *Journal of Policy Studies*, *39*, 53–59.

Sacks, H. (1974). On the analysability of stories by children. In R. Turner (Ed.), *Ethnomethodology: Selected readings* (pp. 216–232). Penguin.

Sacks, H. (1984). On doing 'being ordinary'. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 413–429). Cambridge University Press.

Sacks, H. (1992). *Lectures on conversation* (2 Vols.). Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Sandlund, E., & Sundqvist, P. (2019). Doing versus assessing interaction competence: Contrasting L2 test interaction and teachers' collaborative grading of a paired speaking test. In S. Kunitz & R. Salaberry (Eds.), *Teaching and testing L2 interactional competence: Bridging theory and practice* (pp. 357–396). Routledge.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System, 45*, 79–91. https://doi.org/10.1016/j.system.2014.05.004

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032

Salaberry M. R., Burch A. R. (Eds.). (2021). *Assessing speaking in context: Expanding the construct and its applications*. Multilingual Matters.

Sawaki, Y. (2017). University faculty members' perspectives on English language demands in content courses and a reform of university entrance examinations in Japan: A needs analysis. *Language Testing in Asia, 7*(1), 1–16. https://doi.org/10.1186/s40468-017-0043-2

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361–382.

Schegloff, E. A. (1979). Identification and recognition in telephone conversation openings. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 23–78). Irvington Publishers.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology, 97*(5), 1295–1345. https://doi.org/10.1086/229903

Schegloff, E. A. (1996). Confirming allusions: Toward an empirical account of action. *American Journal of Sociology, 102*(1), 161–216. https://doi.org/10.1086/230911

Schegloff, E. A. (2007a). A tutorial on membership categorization. *Journal of pragmatics, 39*(3), 462–482. https://doi.org/10.1016/j.pragma.2006.07.007

Schegloff, E. A. (2007b). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press.

Schmidt, R. (1983). Interaction, acculturation and the acquisition of communicative competence. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and second language acquisition* (pp. 137–174). Newbury House.

Schönfeldt, J., & Golato, A. (2003). Repair in chats: A conversation analytic approach. *Research on Language and Social Interaction, 36*, 241–284. https://doi.org/10.1207/S15327973RLSI3603_02

Schumann, J. H. (1978). *The pidginization process: A model for second language acquisition* (pp. 367–379). Newbury House Publishers.

Schutz, A. (1962). *Collected papers I: The problem of social reality*. Martinus Nijhoff.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education, 1*, 9–23. https://doi.org/10.1007/BF00143275

Seale, C. (1999). *The quality of qualitative research*. Sage.

Searle, J. R. (1969). *Speech acts*. Cambridge University Press.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, *5*(1), 1–23. https://doi.org/10.1017/S0047404500006837

Searle, J. R. (1992). Conversation. In: H. Parret & J. Verschueren (Eds.), *On Searle on conversation* (pp. 7–30). Benjamins.

Seedhouse, P., & Egbert, M. (2006). *The interactional organization of the IELTS speaking test*. IELTS Research Reports 2006. https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report6.ashx

Seedhouse, P., & Nakatsuhara, F. (2018). *The discourse of the IELTS Speaking Test: Interactional design and practice*. Cambridge University Press.

Serafini, E. J., Lake, J. B., & Long, M. H. (2015). Needs analysis for specialized learner populations: Essential methodological improvements. *English for Specific Purposes*, *40*, 11–26. https://doi.org/10.1016/j.esp.2015.05.002

Sert, O., & Balaman, U. (2018). Orientations to negotiated language and task rules in online L2 interaction. *ReCALL*, 1–20. https://doi.org/10.1017/S0958344017000325

Shohamy, E. (1996). Competence and performance in language testing. In G. Brown, K. Malmknaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 138–151). Cambridge University Press.

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. Wiley-Blackwell.

Siegler, F. A. (1967). Unconscious intentions. *Inquiry, 10*(1–4), 251–267. https://doi.org/10.1080/00201746708601492

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.

Steensig, J., & Drew, P. (2008). Introduction: Questioning and affiliation/disaffiliation in interaction. *Discourse Studies, 10*(1), 5–15. https://doi.org/10.1177%2F1461445607085581

Steensig, J. (2019). Conversation analysis and affiliation and alignment. In C. A. Chapelle (Ed.), *The encyclopaedia of applied linguistics* (pp. 1–6). Blackwell.

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction, 41*(1), 31–57. https://doi.org/10.1080/08351810701691123

Stokoe, E. (2012). Moving forward with membership categorization analysis: Methods for systematic analysis. *Discourse Studies, 14*(3), 277–303. https://doi.org/10.1177%2F1461445612441534

Stubbe, M., Lane, C., Hilder, J., Vine, E., Vine, B., Marra, M., Holmes, J., & Weatherall, A. (2003). Multiple discourse analyses of a workplace interaction. *Discourse Studies*, *5*(3), 351–88. https://doi.org/10.1177%2F1461445603 0053004

Swales, J. (2001). EAP-related linguistic research: An intellectual history. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 42–54). Cambridge University Press.

Taguchi, N. (2012). *Context, individual differences, and pragmatic competence*. Multilingual Matters.

Taguchi, N. (2015). Pragmatics in Chinese as a second/foreign language. *Studies in Chinese Learning and Teaching*, *1*(1), 3–17.

Tai, K. W. H., & Dai, D. W. (2023). Observing a teacher's interactional competence in an ESOL classroom: a translanguaging perspective. *Applied Linguistics Review*. Advance online publication. https://doi.org/10.1515/applirev-2022-0173

ten Have, P. (2007). *Doing conversation analysis*. Sage.

Terhune, N. M. (2016). Language learning going global: Linking teachers and learners via commercial Skype-based CMC. *Computer Assisted Language Learning*, *29*(6), 1071–1089. https://doi.org/10.1080/09588221.2015.1061020

Thurlow, C., & Mroczek, K. (Eds.). (2011). *Digital discourse: Language in the new media*. Oxford University Press.

Timpe-Laughlin, V. (2017). Adult learners' acquisitional patterns in L2 pragmatics: What do we know? *Applied Linguistics Review, 8*(1), 101–129. https://doi.org/10.1515/applirev-2015-2005

Toulmin, S. E. (2003). *The uses of argument (updated ed.)*. Cambridge University Press.

Trace, J. (2016). *A validation argument for cloze test item function in second language assessment*. [Unpublished Ph.D. dissertation, University of Hawai'i at Manoa].

Tsagari, D. (2020). Language assessment literacy: Concepts, challenges and prospects. In S. Hidri. (Ed.), *Perspectives on language assessment literacy: Challenges for improved student learning* (pp. 13–33). Routledge.

Tudini, V. (2010). *Online second language acquisition: Conversation analysis of online chat*. Bloomsbury Publishing.

Tudini, V., & Liddicoat, A. J. (2017). Computer-mediated communication and conversation analysis. In S. L. Thorne & S. May (Eds.), *Language, education and technology* (3rd ed.). Springer International.

Turnbull, W., & Carpendale, J. (2001, Jan). The morality of interaction [paper presentation]. Association for Moral Education conference. Vancouver, Canada. https://www.researchgate.net/publication/291330842_The_morality_of_interaction

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback. The Australian review of applied linguistics series, 13* (pp. 55–79). Cambridge University Press.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49–70. https://doi.org/10.2307/3588360

Turowetz, J. J., & Maynard, D. W. (2010). Morality in the social interactional and discursive world of everyday life. In S. Hitlin & S. Vaisey (Eds.), *Handbook of the Sociology of Morality* (pp. 503–526). Springer.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3–12. https://doi.org/10.1093/elt/49.1.3

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508. https://doi.org/10.2307/3586922

Walker, G. (2013). Phonetics and prosody in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 455–474). Wiley-Blackwell.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, *24*(2), 155–183. https://doi.org/10.1177%2F0265532207076362

Wang, H. (2011). Chinese for business professionals: The workplace needs and business Chinese textbooks. *Global Business Languages*, *16*(1), 27–41.

Waring, H. Z. (2013). 'How was your weekend?': Developing the interactional competence in managing routine inquiries. *Language Awareness, 22*(1), 1–16. https://doi.org/10.1080/09658416.2011.644797

Watanabe, A. (2017). Developing L2 interactional competence: Increasing participation through self-selection in post-expansion sequences. *Classroom Discourse, 8*(3), 271–293. https://doi.org/10.1080/19463014.2017.1354310

Watson, D. R. (1978). Categorization, authorisation and blame-negotiation in conversation. *Sociology, 12*(1), 105–13. https://doi.org/10.1177%2F003803857801200106

Wilson, D., & Sperber, D. (1981). On Grice's theory of conversation. In P. Werth (Ed.), *Conversation and discourse* (pp. 155–78). Croom Helm.

Winsteps. (n.d., a). *Table 3.1 Summaries of persons and items*. https://www.winsteps.com/winman/table3_1.htm

Winsteps. (n.d., b). *Inter-rater and intra-rater Reliability*. https://www.winsteps.com/facetman/inter-rater-reliability.htm

Wittgenstein, L. (1953). *Philosophical investigations*. (G. E. M. Anscombe, Trans). Basil Blackwell.

Wolfson, N. (1990). The bulge: A theory of speech behavior and social distance. *Penn Working Papers in Educational Linguistics, 2*(1), 55–83. https://repository.upenn.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1009&context=wpel

Wong, J., & Waring, H. Z. (2020). *Conversation analysis and second language pedagogy: A guide for ESL/EFL teachers* (2nd ed.). Routledge.

Wright, B. D. (1992). IRT in the 1990s: Which models work best. *Rasch measurement transactions, 6*(1), 196–200.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509–511.

Wu, R. J. (2003). *Stance in talk: A conversation analysis of Mandarin final particles.* John Benjamins.

Wu, R. J. R. (2016). Doing conversation analysis in Mandarin Chinese. *Chinese Language and Discourse*, *7*(2), 179–209. https://doi.org/10.1075/cld.7.2.01wu

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods.* [Unpublished Ph.D. dissertation, University of Hawai'i at Manoa].

Youn, S. J. (2014). Measuring syntactic complexity in L2 pragmatic production: Investigating relationships among pragmatics, grammar, and proficiency. *System*, *42*, 270–287. https://doi.org/10.1016/j.system.2013.12.008

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225. https://doi.org/10.1177%2F0265532214557113

Youn, S. J. (2018). Task-based needs analysis of L2 pragmatics in an EAP context. *Journal of English for Academic Purposes, 36*, 86–98. https://doi.org/10.1016/j.jeap.2018.10.005

Youn, S. J., & Burch, A. R. (Eds.) (2020). Where Conversation Analysis meets language assessment. *Papers in Language Testing and Assessment, 9*(1).

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning, Vol. 2* (pp. 426–443). Routledge.

Zhang, Y. (2016). *Development of second language interactional competence: Agreement and disagreement negotiation by learners of Mandarin*. [Unpublished Ph.D. dissertation, University of Melbourne].

Zhao, J. (2008). Concept and mode of teaching Chinese as a second language. *Chinese Teaching in the World, 1*, 93–107.

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586. https://doi.org/10.1017/S027226311500025X

Appendix I: S-H interview protocol

Lead-in

As you know I'm interested in the challenges you have experienced when using Chinese to communicate with other Chinese-speaking people in China[15]. What were the situations or what was it about interaction in Chinese you find **most** tricky, uncomfortable, or difficult to handle? We should focus on the situations where you use Chinese instead of other languages. Are there one or two situations that stand out for you that you might want to talk about?

Sample prompting questions for L2-Chinese speakers based on Dinkins (2005) and Roulston (2010).

- What's your definition of interaction?
- In your opinion, what was the **most** challenging situation for interaction in Chinese, or what makes interaction in Chinese **most** challenging?
- Tell me about …
- You mentioned that you had _____; could you tell me more about _____
- You mentioned when you were doing _____, _____ happened. Could you give me a specific example of that?
- Thinking back to that time, what was that like for you?
- You mentioned earlier that you _____. Could you describe in detail what happened?
- You mentioned _____. I can give you an example that might be similar to what you were saying _____. Can you think of something similar?
- You mentioned _____ but I think it contradicts what you said earlier _____. What do you think?
- Let's revisit our definition of interaction in Chinese. Do you still think _____?

15 The questions listed here are for the L2-Chinese speaker group. The questions were adapted for the L1-Chinese interactant and Chinese teacher groups, as discussed in Section 4.2.2.1

Appendix II: Norming questionnaire

English translation

Norming item 1

Work P+ 2nd PP voice message – Criticising

You are the team leader for a project in a company. One of your team members is Wu Han, who is the son of the company's general manager. You and Wu Han are similar in age. Wu Han has been showing up to work late frequently and is taking forever to hand back a report you asked him to write. His behaviour has caused a negative impact on the team morale and has eroded your authority as the team leader. This morning Wu Han didn't come to work and sent you a voice message at midday. After listening to it you decide to reply with a voice message.

Wu Han's voice message: *Team leader, sorry I woke up late today so I probably won't be coming to work. The report you mentioned the other day is too hard for me. How about you find someone else to write it? Thank you.*

Q1: If you were to reply to Wu Han's voice message, what would you say?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to reply with a voice message?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to comment on Wu Han's behaviour in your voice message?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Wu Han?

0	1	2	3	4	5	6	7	8	9	10
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Wu Han's ranking in the company in comparison to yours?

0	1	2	3	4	5	6	7	8	9	10
Much lower			Lower		Same		Higher			Much higher

Norming item 2

Life P= 2nd PP voice message-Breaking bad news

Wang Si is your friend. You two are similar in age. She went traveling overseas for three months and asked you to look after her dog, Dou Dou, whom she had had for eight years. Last week you took Dou Dou out for a stroll, met a friend and got into a chat. You were distracted by the chat and let go of the leash. When you realised it Dou Dou was already gone. You have looked for Dou Dou for three days but still haven't found him. Just now Wang Si sent you a voice message. After listening to it you decide to reply with a voice message.

Wang Si's voice message: *Hey, buddy, how are you doing? Will you be home this afternoon? I just realised Dou Dou is due for his check-up today and I have arranged for a friend to pick him up this afternoon and take him to the vets. Could you give Dou Dou to my friend when he arrives? Thanks!*

Q1: If you were to reply to Wang Si's voice message, what would you say?

Q2: Do you find this scenario realistic?

0	1	2	3	4	5	6	7	8	9	10
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to reply with a voice message?

0	1	2	3	4	5	6	7	8	9	10
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to discuss Dou Dou in your voice message?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Wang Si?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Wang Si's seniority in society in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 3

Study P+ 1st PP voice message-Negative commenting

You are a senior college student and Liu Hong is a new student in your department. You are aware that Liu Hong has been under a lot of stress with study and is on treatment for mild depression. This morning Liu Hong asked you to have a read at one of her final essays and offer her some feedback. You read the essay and noticed it was badly written, with chunks of paragraphs plagiarised from other sources. Liu Hong told you she had to submit this paper tomorrow morning. You decide to send her a voice message and talk about her essay.

Q1: If you were to send Liu Hong a voice message, what would you say?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to send a voice message to discuss the issue on hand?

0	1	2	3	4	5	6	7	8	9	10
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to discuss Liu Hong's essay in your voice message?

0	1	2	3	4	5	6	7	8	9	10
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Liu Hong?

0	1	2	3	4	5	6	7	8	9	10
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Liu Hong's institutional ranking at school in comparison to yours?

0	1	2	3	4	5	6	7	8	9	10
Much lower			Lower		Same		Higher			Much higher

Norming item 4

Life P+ video chat -Criticising

You recently went to a different city for work for six months and sub-let the apartment you rented to your best friend's son Wang Bin. Your friend is very grateful to you for that. You used to visit your friend a lot and always had a good impression of Wang Bin. Today your neighbour texts you, telling you that lately there has been a lot of noise and loud music in your apartment late at night. Sometimes the neighbours also see drunken youths coming in and out of your apartment. You want to discuss this with Wang Bin but since you are still away in a different city so you decide to talk to him via WeChat video chat.

Q1: If you were to have a video chat with Wang Bin, what would you talk about?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to communicate with Wang Bin via video chat?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to talk about Wang Bin's behaviour in the video chat?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Wang Bin?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Wang Bin's seniority in society in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 5

Study P= video chat – Disagreeing

You and Li Ran are college students and have been roommates for a few years. Summer is approaching and Li Ran suggested that you two chip in to buy an air-conditioner together. Though you live with Li Ran, you have never told him/

her that your financial situation is not that good and you don't have the money for an air-conditioner. Besides you are worried that once you two got the air-conditioner the electricity bills would sky-rocket, which you wouldn't be able to afford. You are not living in the dorm at the moment and Li Ran said he/she wanted to have a chat with you. You agreed to have a video chat with him/her.

Q1: What would you say if Li Ran brought up the air-conditioner during the video chat again?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to discuss the issue on hand via video chat?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to talk about the air-conditioner with Li Ran?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Li Ran?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Li Ran's institutional ranking at school in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 6

Life P- 1st PP voice message – Requesting

Zhao Ran is your close friend's mother. She currently lives overseas and runs a Taobao Daigou shop (shops that buy stuff from overseas and send them to China). You met Zhao Ran a few times when she was in China and she was very nice to you. Recently you bought three expensive designer bags from her shop. However when you received them you noticed the quality was shoddy. You took the bags to the retail stores and were told they were knockoffs. According to Taobao policies you have the right to request a full refund in these situations. You decide to send Zhao Ran a voice message to discuss this issue.

Q1: If you were to send Zhao Ran a voice message, what would you say?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to send a voice message to discuss the issue on hand?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to discuss the bags with Zhao Ran in your voice message?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Zhao Ran?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Zhao Ran's seniority in society in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 7

Work P= 1st PP voice message – Cancelling

You work for a company and there are 18 people in your department. Zhang Bin is your colleague and he organised a team-bonding session for the department this Saturday morning. Everyone in the department, including the managers, said they would go and you told Zhang Bin that you would go as well. Lately you are looking for a different job and your potential employer is interested in you but she can only interview you this Saturday morning. If you went to the interview you would miss out on the team-bonding session organised by Zhang Bin. Now you decide to send Zhang Bin a voice message telling him your situation for this Saturday.

Q1: If you were to send Zhang Bin a voice message, what would you say?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to send a voice message to discuss the issue on hand?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to discuss the issue with Zhang Bin in your voice message?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Zhang Bin?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Wu Han's ranking in the company in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 8

Work P- Video chat – Complaining

You and Tang Li work in the same company. You two are close in age, work in similar areas and you have always worked harder than her. Tang Li has been asking for leave all the time for reasons such as travelling, study and visiting doctors. Your manager Li Jia has always granted Tang Li her leave. Last week your boyfriend/girlfriend needed to go to the hospital for an operation. You wanted to take a day off to take him/her to the hospital but Li Jia rejected your request. In the end your boyfriend/girlfriend went to the operation accompanied by a friend instead of you. You thought Li Jia was being unfair in terms of your leave request and you told her you wanted to have a chat with her. Li Jia said she was working from home lately but she agreed to have a video chat with you.
Q1: If you were to have a video chat with Li Jia, what would you talk about?
Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to communicate with Li Jia via video chat?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to talk about the leave in the video chat?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Li Jia?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Li Jia's ranking in the company in comparison to yours?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Much lower			Lower		Same		Higher			Much higher

Norming item 9

Study P- 2nd PP voice message – Refusing

You are a college student and both Chen Gang and Wu Hong are teachers in your school. You have taken classes from both teachers but you are more interested in Wu Hong's research areas. The summer break is approaching and you already applied for an internship with Wu Hong. Just now you received a voice message from Li Gang. After listening to it you decide to send a voice message back in reply.

Chen Gang's voice message: Hey there is something I wanted to talk to you about. You know the summer break is coming. If you don't have any plans would you be interested in doing an internship in my research group? I think you are very competent and hardworking. Having some internship experience under your belt will be good for your study and work in the future. What do you think?

Q1: If you were to reply to Chen Gang's voice message, what would you say?

Q2: Do you find this scenario realistic?

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Very unrealistic			Not realistic		Average		Realistic			Very realistic

Q3: Based on the scenario in this story, do you find it appropriate to reply with a voice message?

0	1	2	3	4	5	6	7	8	9	10
Very inappropriate			Inappropriate		Average		Appropriate			Very appropriate

Q4: If you were the protagonist, would you find it difficult to reply to Chen Gang in your voice message?

0	1	2	3	4	5	6	7	8	9	10
Very easy			Easy		Average		Difficult			Very difficult

Q5: If you were the protagonist, how familiar would you say you were with Chen Gang?

0	1	2	3	4	5	6	7	8	9	10
Very unfamiliar			Unfamiliar		Average		Familiar			Very familiar

Q6: If you were the protagonist in this story, how would you rate Chen Gang's institutional ranking at school in comparison to yours?

0	1	2	3	4	5	6	7	8	9	10
Much lower			Lower		Same		Higher			Much higher

Chinese version[16]

人际关系认可度问卷调查
亲爱的朋友，你好！
本问卷调查中你会看到九个情景，每个情景下有六个小问题。题干中有的文字有背景颜色，这仅仅是为了强调一些可能会被忽视的信息。

16 Here only the first-round norming questionnaire is included. The second-round questionnaire was revised based on L1-Chinese respondents' responses in the first-round questionnaire, as explained in Sections 4.2.3.2 and 4.2.4.2. Please also note the order of the items in the norming questionnaire is different from the order in the IC test in Appendix III.

如果问题是问答题，如“假如你给A发语音，你大概会说些什么？”，请简要地写一下你可能会谈及的内容。
如果问题是打分题，请把你觉得适合的分数写在题干边的括号内。如：0代表“非常不真实”，10代表“非常真实”。0到10是一个由“非常不真实”到“非常真实”的过渡，如果你觉得情景的真实性介于“真实”（7）和“非常真实”（10）之间，请根据你的判断选择8或者9，9比8更“真实”一点，其它问题以此类推。

例样：

你觉得这个情景真实吗？（**8**）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不真实			不真实		一般		真实			非常真实

请根据你所了解的情况以及真实感受进行回答，完成本次问卷调查大约需要10分钟。

谢谢你的参与。

戴维

问卷第一题
你是公司一个项目的负责人，管理一个团队，团队里有你上司的侄子吴翰，你和吴翰年龄差不多。吴翰最近上班一直迟到，你布置任务叫他写一篇报告，他一直没交。吴翰的工作态度给整个团队带来了很不好的影响。今天吴翰没来上班，中午他给你发了一条微信语音，你听后决定回复吴翰。
吴翰的语音："组长，我今天起来晚了，要么我就在家办公，行吗？您上次说的那个报告，我实在是不会写，您要么找别人写，好不好？谢谢！"

Q1: 假如你来回复吴翰的微信语音，你会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得现实生活中用微信语音来回复吴翰合适吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不合适			不合适		一般		合适			非常合适

Q4: 假如你是故事主人公，要来点评一下吴翰的行为，你觉得难开口吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常容易			容易		一般		难			非常难

Q5: 假如你是故事主人公，你觉得你和吴翰的熟悉程度是？（）

0	1	2	3	4	5	6	7	8	9	10
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事主人公，你觉得吴翰在公司的级别和你相比 （）

0	1	2	3	4	5	6	7	8	9	10
低很多			低		一样		高			高很多

问卷第二题
王思是你的朋友，你们年龄相近。王思出国旅游三个月，拜托你照顾她养了8年的宠物狗——豆豆。你上周带豆豆出去散步时因为碰到熟人聊天一走神松了牵豆豆的绳子，等你想起来时“豆豆”已经跑走不见了。你找了一个星期也没找到“豆豆”。今天王思给你发了条微信语音，你听后决定用语音回复王思，告诉她“豆豆”的事。
王思语音：“嗨，亲，你最近还好吗，住得还习惯吗？下午会在家吧？我刚想起来今天‘豆豆’要体检了，我让我一个朋友下午来接‘豆豆’，带它去宠物医院。等我朋友来时麻烦你把‘豆豆’交给他好吗？谢谢啦！”

Q1: 假如你来语音回复王思，你会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得现实生活中听了王思的语音后用语音回复合适吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不合适			不合适		一般		合适			非常合适

Q4: 假如你是故事主人公，要和王思说“豆豆”的事，你觉得难开口吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常容易			容易		一般		难			非常难

Q5: 假如你是故事主人公，你觉得你和王思的熟悉程度是？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事主人公，你觉得王思的辈分和你相比（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
低很多			低		一样		高			高很多

问卷第三题
你是大学高年级学生，刘红是你的学妹，你知道她最近因为学习压力大得了抑郁症，在接受治疗。今天早上刘红请你帮她看一下她的期末论文。你看了以后发现她写得很糟糕，语句不通，逻辑混乱，还有很多地方是直接从别的文章里抄来的。刘红说她明天早上要交这篇论文，你决定给刘红发一条微信语音，说一下她论文的事。

Q1: 假如你来给刘红发语音，你会说些什么？

答：。

Q2: 你觉得这个情景真实吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得现实生活中用微信语音来说这件事合适吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不合适			不合适		一般		合适			非常合适

Q4: 假如你是故事主人公，要和刘红说她文章的事，你觉得难开口吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常容易			容易		一般		难			非常难

Q5: 假如你是故事主人公，你觉得你和刘红的熟悉程度是？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事主人公，你觉得刘红在学校里的资历和你相比（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
低很多			低		一样		高			高很多

问卷第四题
最近你去外地出差半年，你把租的公寓转租给了你最好的朋友的儿子王斌，你的好朋友为此非常感激你。你以前常去好朋友家做客，对王斌印象很好。今天你公寓的邻居给你发短信，说你家这两天深夜时有很响的

音乐声和吵闹的声音，有时还看到醉醺醺的年轻人进进出出。你想和王斌谈一下这件事，由于你还在出差，你决定用微信视频来和他交流。

Q1: 假如你和王斌微信视频，你大概会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得在本题的情景设定中，用微信视频交流合理吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不合理			不合理		一般		合理			非常合理

Q4: 假如你是故事的主人公，要和王斌说他的行为，你觉得有难度吗？()

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常容易			容易		一般		难			非常难

Q5: 假如你是故事的主人公，你觉得你和王斌的熟悉程度是？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事的主人公，你觉得王斌的辈分和你相比（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
低很多			低		一样		高			高很多

问卷第五题

你和李然是大学同学，并且是室友。夏天快到了，李然建议一起出钱买个空调。你经济条件不是很好，没有钱买空调，同时你也觉得夏天没那么热，用电风扇就可以了。你这两天不在学校，李然还想和你讨论一下空调的事，你答应和李然视频。

Q1: 假如李然视频时再次提到想一起买空调，你会怎么回复？

答：。

Q2: 你觉得这个情景真实吗？（）

0 1 2 3 4 5 6 7 8 9 10
非常不真实 不真实 一般 真实 非常真实

Q3: 你觉得现实生活中用视频讨论这件事合适吗？（）

0 1 2 3 4 5 6 7 8 9 10
非常不合适 不合适 一般 合适 非常合适

Q4: 假如你是故事主人公，要和李然说空调一事，你觉得难开口吗？（）

0 1 2 3 4 5 6 7 8 9 10
非常容易 容易 一般 难 非常难

Q5: 假如你是故事主人公，你觉得你和李然的熟悉程度是？（）

0 1 2 3 4 5 6 7 8 9 10
非常不熟 不熟 一般 熟 非常熟

Q6: 假如你是故事主人公，你觉得李然在学校里的资历和你相比（）

0 1 2 3 4 5 6 7 8 9 10
低很多 低 一样 高 高很多

问卷第六题
6. 赵冉是你好朋友的妈妈，她目前定居国外并开了一家淘宝代购店。赵冉以前在国内时你见过她几次，她一直对你很友好。你最近在赵冉店里买了三个很昂贵的名牌包。你收到时觉得包的质量不好，拿到专卖店里去询问，店员说都是假货。按照淘宝的政策，在这种情况下你有充分理由申请退款 。你决定给赵冉发一条微信语音说一下这件事。

Q1: 假如你来给赵冉发微信语音，你会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0 1 2 3 4 5 6 7 8 9 10
非常不真实 不真实 一般 真实 非常真实

Q3: 你觉得在本题的情景设定中，用微信语音交流合理吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不合理			不合理		一般		合理			非常合理

Q4: 假如你是故事的主人公，要和赵冉说“包包”一事，你觉得有难度吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常容易			容易		一般		难			非常难

Q5: 假如你是故事的主人公，你觉得你和赵冉的熟悉程度是？（）

0	1	2	3	4	5	6	7	8	9	10
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事的主人公，你觉得赵冉的辈分和你相比 （）

0	1	2	3	4	5	6	7	8	9	10
低很多			低		一样		高			高很多

问卷第七题

你在一家公司工作，你们部门一共有18个人。张斌是你的同事，他这周六上午给整个部门组织一个团队建设的活动。部门所有人，包括领导都说会去，你也答应张斌说你会去。你最近计划跳槽，新公司的面试官对你很有兴趣，但他只有周六上午有时间来面试你。你去参加面试的话就会错过张斌组织的团建活动。你现在要给张斌发一条微信语音说一下周六的事情。

Q1: 假如你来给张斌发微信语音，你会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得在本题的情景设定中，用微信语音交流合理吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不合理			不合理		一般		合理			非常合理

Q4: 假如你是故事的主人公，要和张斌说周六的事，你觉得有难度吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常容易			容易		一般		难			非常难

Q5: 假如你是故事的主人公，你觉得你和张斌的熟悉程度是？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事的主人公，你觉得张斌在公司的级别和你相比（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
低很多			低		一样		高			高很多

问卷第八题

你和唐丽在一家公司工作，年龄差不多，工作的内容也一样，你工作一直比唐丽认真。唐丽经常因为各种原因请假，例如：旅游、进修、看医生等，而你们的经理李佳也总是准唐丽的假。你恋人上周要做手术，你想请一天假陪恋人去医院，可李佳却没有同意，最终你的恋人只能找朋友陪着去医院。你觉得李佳在请假这件事上对你不公平，你和李佳说想沟通一下，李佳说她最近在家办公，不去公司，不过可以和你微信视频谈。

Q1: 假如你和李佳微信视频，你大概会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得在本题的情景设定中，用微信视频交流合理吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不合理			不合理		一般		合理			非常合理

Q4: 假如你是故事的主人公，要和李佳谈请假这件事，你觉得有难度吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常容易			容易		一般		难			非常难

Q5: 假如你是故事的主人公，你觉得你和李佳的熟悉程度是？（）

0	1	2	3	4	5	6	7	8	9	10
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事的主人公，你觉得李佳在公司的级别和你相比 （）

0	1	2	3	4	5	6	7	8	9	10
低很多			低		一样		高			高很多

问卷第九题
你是大学生，陈刚和吴虹是你们专业的老师，两位老师的课你都上过，不过你对吴虹的课题研究更感兴趣。暑假快到了，你决定申请跟吴虹做实习。现在你收到陈刚的一条微信语音，你听后决定回复陈刚。
陈刚的语音："老师有个事要和你说一下，暑假快到了，你要是没什么安排的话，想不想来我的课题组实习啊？我觉得你能力很强，也很聪明好学，有实习经历对你以后学习工作都会有帮助的。你觉得怎么样啊？"

Q1: 假如你来回复陈刚的语音，你会说些什么？

答：

Q2: 你觉得这个情景真实吗？（）

0	1	2	3	4	5	6	7	8	9	10
非常不真实			不真实		一般		真实			非常真实

Q3: 你觉得在现实生活中，听了陈刚的语音后，用语音回复合适吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不合适			不合适		一般		合适			非常合适

Q4: 假如你是故事主人公，要回复陈刚的语音，你觉得难开口吗？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常容易			容易		一般		难			非常难

Q5: 假如你是故事主人公，你觉得你和陈刚的熟悉程度是？（）

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
非常不熟			不熟		一般		熟			非常熟

Q6: 假如你是故事主人公，你觉得陈刚在学校里的资历和你相比（）

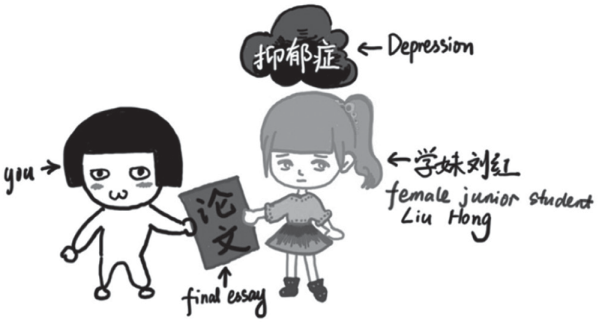
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
低很多			低		一样		高			高很多

Appendix III: The IC test

Item 1_study_1st pp_P +_negative commenting

你是大学高年级学生，刘红是你的学妹，你知道她最近因为学习压力大，她生病了，有了抑郁症，常常需要吃药。现在到期末了，刘红请你帮她看一下她的期末论文。你看了刘红的论文后发现她写得很不好，很多地方是从别的文章里抄来的。刘红说她明天早上要交这篇论文，你决定给刘红发语音短信，和她说一下这件事。

You are a senior college student. Liu Hong is a new student in your department. You are aware that Liu Hong has been under a lot of stress with study and is on treatment for depression. This morning Liu Hong asked you to have a read at one of her final essays and offer her some feedback. You read the essay and noticed it was badly written, with chunks of paragraphs plagiarised from other sources. Liu Hong told you she had to submit this paper tomorrow morning. You decide to send her a voice message to talk about her essay.



Item 2_life_1st pp_P-_requesting

你有一个朋友叫张星，张星的妈妈李静住在国外，她在国外开了一家淘宝店。李静在国内的时候你见过她几次，她对你很好。你最近在李静的淘宝店里买了三个很贵的包。你收到的时候，觉得包有问题，所以你拿到专卖店里去问，店员说都是假货。按照淘宝的规定，你可以要求李静把钱还给你。你决定给李静发语音短信说一下这件事。

Zhang Xing is your friend. Zhang Xin's mother, Li Jing, lives overseas and runs a Taobao shop (shops that buy stuff overseas and send them to China). You met Li Jing a few times when she was in China and she was very nice to you. Recently you bought three expensive designer bags from her shop. However when you received them you noticed the quality was shoddy. You took the bags to the retail store and were told they were knockoffs. According to Taobao policies you have the right to request a refund in this situation. You decide to send Li Jing a voice message to discuss this issue.


你 you
朋友张星
friend Zhang Xing
张星妈妈李静
Zhang Xing's mom Li Jing
淘宝网
Taobao.com
an online shop
knock-off
假

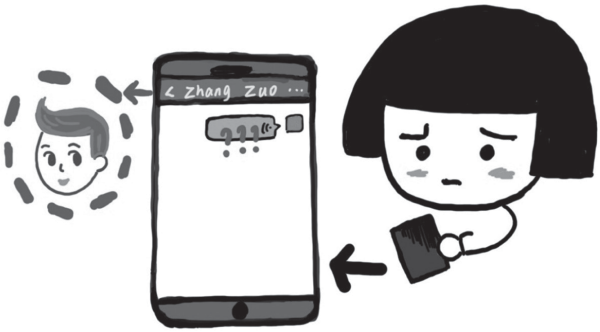
Item 3_work_1st pp_P=_cancelling

你在一家公司工作，张左是你的同事，他这个周六上午会给整个公司安排一个和公司工作有关的团队建设活动。公司里面的所有员工和领导都说会参加，你也和张左说你会参加活动。你最近计划换工作，新公司的经理很喜欢你，可他刚刚告诉你他只有本周六上午有时间面试你。如果你参加面试的话，你就不能参加张左安排的活动。你现在决定给张左发语音短信，说一下周六的情况。

You work for a company and Zhang Zuo is your colleague, who has organised a team-bonding session for the department this Saturday morning. Everyone in the department, including the managers, said they would go, and you told Zhang Zuo that you would go as well. Lately you are looking for a different job and your potential employer is interested in you, but he can only interview you this Saturday morning. If you went to the interview you would miss out on the team-bonding session organised by Zhang Zuo. Now you decide to send Zhang Zuo a voice message telling him your situation for this Saturday.


team bonding activity
this Sat
团队建设
这周六
同事张左
colleague Zhang Zuo
你, you
领导 managers

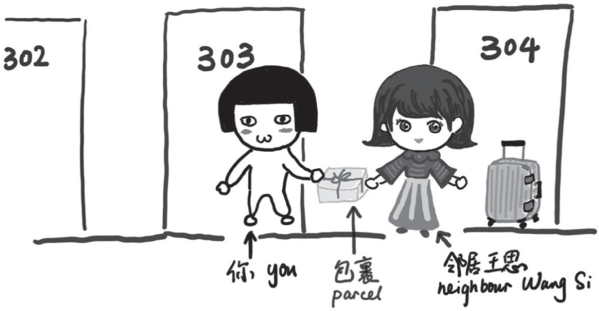

周六新公司面试
new company job
interview this Sat



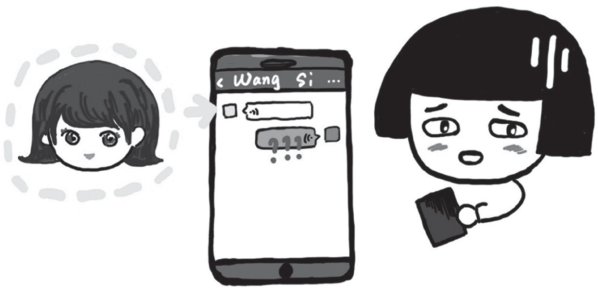
Item 4_life_2nd pp_P=_negative news telling

王思是你的邻居，你们年龄差不多。一周前她要去国外，她离开之前请你帮她寄一个包裹，你答应她你那天就会把包裹寄给她的朋友。可是你后来一个星期里工作比较忙，完全忘记了寄包裹这件事。今天王思给你发了一条语音短信，你听后决定给王思发语音短信说一下这件事。
王思语音："嗨，亲，想问一下那个包裹你寄出去了吗？是这样的，包裹里是我买给一个朋友的生日礼物。我朋友昨天生日，可他说没收到我的礼物。我觉得很奇怪，这都一个礼拜了应该寄到了啊。"

Wang Si is your neighbour. You two are similar in age. One week ago she went to another city for some urgent errands and before she left, she asked you to send a parcel for her. You promised her you would send it on that day. However, you were very busy with work for the following week and completely forgot about the parcel. Just now Wang Si sent you a voice message. After listening to it you decide to reply with a voice message.
Wang Si's message: "Hey, I just wanted to check if you had already sent off my parcel? Actually it is my friend's birthday present in the parcel. My friend had their birthday yesterday and strangely they said they didn't receive my gift. I found it quite weird because the parcel should have arrived after one week."


302
303
304
你 you
包裹 parcel
邻居王思 neighbour Wang Si

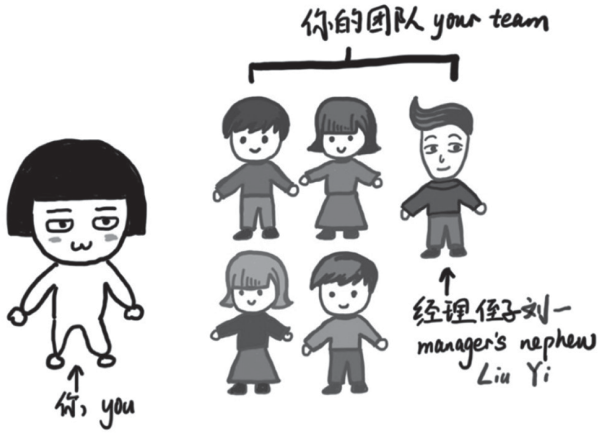

一个星期 for one week ...

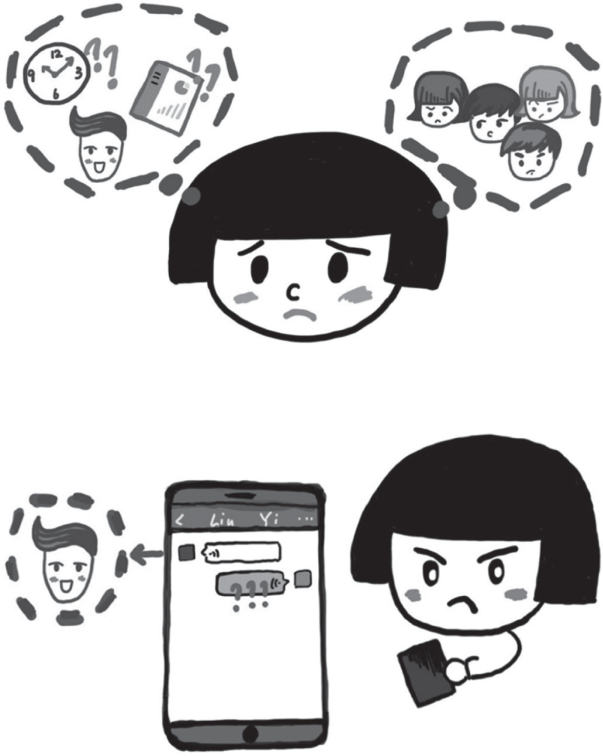

Wang Si ...

Item 5_work_2nd pp_P+_criticising

你是公司一个活动的负责人，管理一个团队，团队里有你经理的侄子刘一。刘一最近上班一直迟到，你叫他写一篇报告，他也一直没交给你。刘一的工作态度给整个团队带来了很不好的影响，他的行为也让你很难管理这个团队。今天早上刘一没有来上班，中午时他给你发了一条语音短信，你听了以后决定给刘一发语音短信，说一下这件事。
刘一的语音："组长，我今天起来晚了，就不来上班了，可以吗？你上次说的那个报告，我实在是不会写，你找别人写，好不好？谢谢！"

You are the team leader for a project in a company. One of your team members is Liu Yi, who is the nephew of your manager. Liu Yi has been showing up to work late frequently and is taking forever to hand back a report you asked him to write. His behaviour has caused a negative impact on the team morale and has eroded your authority as the team leader. This morning Liu Yi didn't come to work and sent you a voice message at midday. After listening to it you decide to reply with a voice message.
Liu Yi voice message: "Hi team leader, sorry I got up late today. How about I just take today off? The report you asked me to write is too hard for me. Can you find someone else to write it? Thanks!"





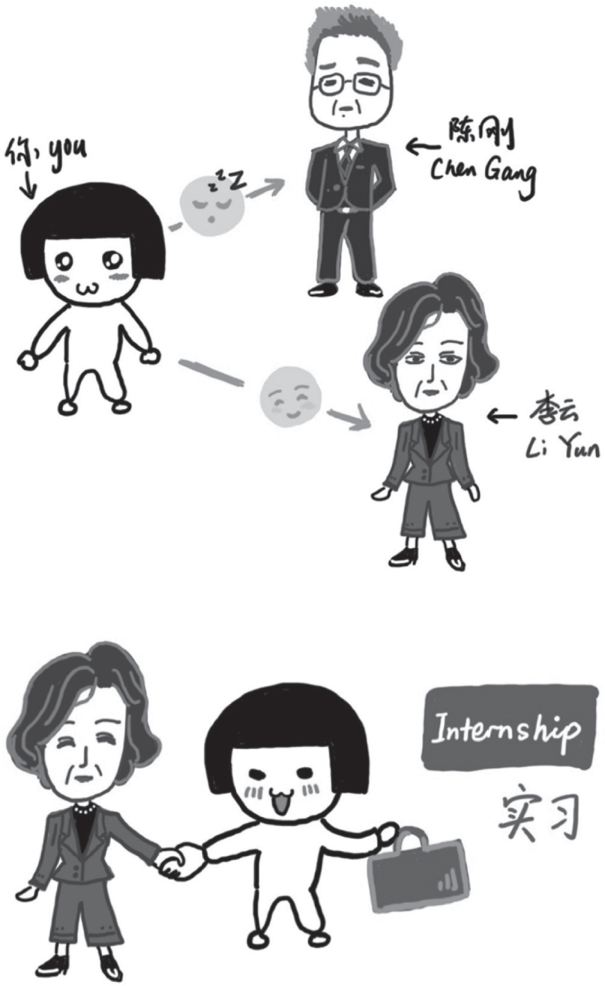
Item 6_study_2nd pp_P-_refusing

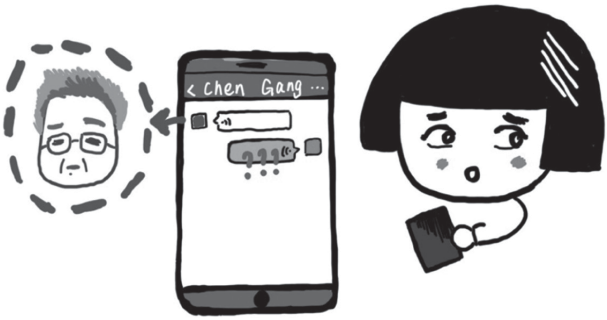
你是大学生，陈刚和李云是你们的老师　。两位老师的课你都上过，不过你对李云的课更感兴趣。暑假快到了，你已经和李云说了你想要去她那里实习，李云也答应你了。但是现在你收到陈刚的一条语音短信，你听后决定给陈刚发语音，说一下这件事。
陈刚的语音："嗨，同学，老师有个事要和你说一下，暑假快到了，你要是没什么安排的话，想不想来我的课题组实习啊？我觉得你能力很强，也很聪明好学。再说有实习经历对你以后学习工作都会有帮助的。你觉得怎么样啊？"

You are a college student. Both Chen Gang and Li Yun are teachers in your school. You have taken classes from both teachers but you are more interested in Li Yun's research areas. The summer break is approaching and you have already

applied for an internship with Li Yun and Li Yun accepted you. Just now you received a voice message from Chen Gang. After listening to it you decide to send a voice message back in reply.

Chen Gang's message: "Hey, there is something I'd like to discuss with you. Summer break is approaching. If you don't have any other plans, would you be interested in an internship in my team? I think you are very capable, smart and inquisitive. Having internship experience under your belt will also be good for your study and work in the future. What do you think?"





Item 7_life_video_P+_criticising

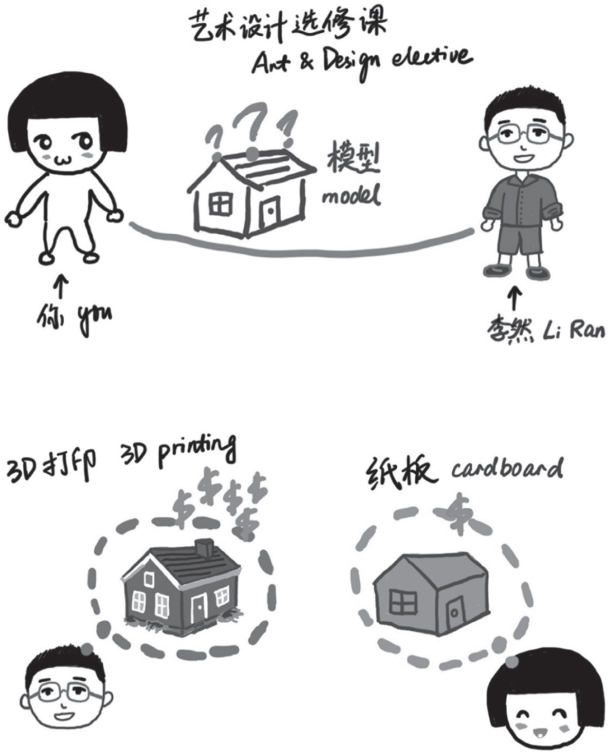
你最近要去国外出差半年，你把你的房子租给王超的儿子王冰。王超是你最好的朋友，他经常帮助你。你以前去王超家的时候也见过王冰，你对他的印象很好。今天你房子的管理员给你打电话，说你家最近深夜里经常有很大的音乐声，有时侯还看到喝醉的年轻人，邻居们都很不高兴。你想和王冰谈一下这件事，因为你还在出差，你决定和他视频交流一下。

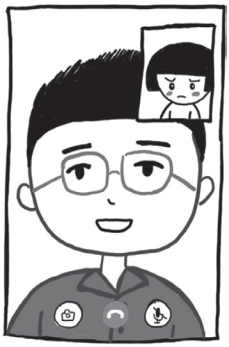
You recently went to a different city for work for six months and sub-let the apartment you rented to Wang Chao's son Wang Bin. Wang Chao is your best friend and has helped you a lot over the years. You also met Wang Bin before when you visited Wang Chao and had a good impression of him. Today the building manager of your apartment called you, telling you that recently there had been a lot of noise and loud music in your apartment late at night. Sometimes the neighbours also saw drunken youths coming in and out of your apartment. You want to discuss this with Wang Bin but since you are still away so you decide to talk to him via video chat.

Item 8_study_video_P=_disagreeing

你是大学生，李然是你同年级的同学，这学期你们选了一门艺术设计的选修课，这门课的期末作业是做一个房子的模型，老师让你和李然一起完成这个作业。李然想用3D打印来做这个模型，这样模型会更好看，当然价格也会更高。你没有告诉李然你家里很穷，你没有钱做3D打印，所以你想用便宜的纸板来做模型。你这两天不在学校，李然说有事想和你讨论一下，你答应和李然视频聊天。

You and Li Ran are college students and are in the same year. Both of you take an arts and design elective this semester and its final assignment is to make a house model. You and Li Ran are assigned to work together in a team. Li Ran suggests that you two use 3D printing to make the model since the results will look better, though the cost will also be higher. You haven't told Li Ran that your financial situation is not that good and you are very certain you are unable to shoulder the cost of 3D printing. Therefore, you prefer to use cardboard to build the house because it is cheaper. You are not on campus these days and Li Ran says he wants to have a chat with you. You agree to have a video chat with him.

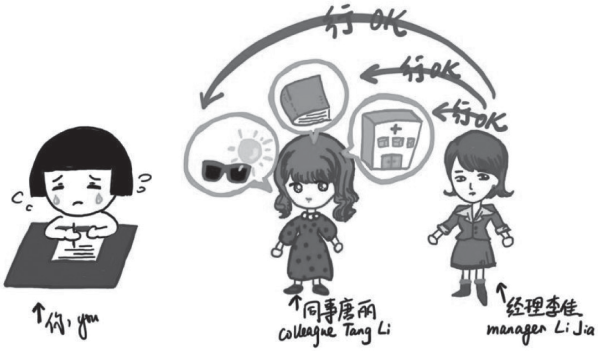




Item 9_work_video_P-_complaining

你和唐丽在一家公司工作，年龄差不多，工作的内容也一样。唐丽今年经常因为各种原因请假，比如：旅游、学习、看医生等，而你们的经理李佳也总是同意唐丽请假。有时候因为唐丽请假，你必须加班把她的工作也完成。你上周想周五请一天假陪恋人去旅游，可是李佳没有同意。你觉得李佳在请假这件事上对你不公平，你很不开心。你想要和李佳说一下请假不公平这件事，李佳最近在家工作，不去公司，不过她同意和你视频聊天。

You and Tang Li work in the same company. You two are close in age and work in similar areas. Tang Li has been asking for leave all the time for reasons such as travelling, study and visiting doctors. Your manager Li Jia has always granted Tang Li her leave. Sometimes because Tang Li was away you had to stay back and finish her work. You wanted to take last Friday off to go on a vacation with your partner but Li Jia rejected your request. You are upset about it as you think Li Jia was being unfair. You told Li Jia you wanted to talk about it. Li Jia says she is working from home lately but she agrees to have a video chat with you.


行 OK
行 OK
行 OK
你, you
同事唐丽
colleague Tang Li
经理李佳
manager Li Jia


不行 No
和恋人旅行
travel with your partner

Appendix IV: The IC rating scale

English version

Disaffiliation control[17]

Band 5 Exemplary	• Disaffiliation is successfully and skillfully remediated, social solidarity unaffected. Disaffiliative actions are either unstated or approached exceedingly strategically in a way that does not cause affront (DAA). • Turn design conforms clearly with dispreferred formats, drawing on a wide range of lexical devices, morphosyntactic features and evidential markers (LDC). • Excellent control of phonetic and prosodic features in mitigating disaffiliative stances (PDC).
Band 4 Good	• Disaffiliation is well managed. Strong disaffiliative actions are covertly realized. Weak disaffiliative actions can be explicitly delivered but are admissible in the local context (DAA). • Both lexical and non-lexical devices are recruited to mitigate disaffiliation. Infelicitous choices are very occasional and not interactionally disruptive (LDC). • Phonetic and prosodic features are in tune with turn design in moderating disaffiliation (PDC).
Band 3 Average	• Disaffiliation is hearable though social solidarity is still intact. There is inconsistency in the speaker's management of disaffiliative activities (DAA). • Some conventionalized verbal resources are involved to soften disaffiliation to an acceptable level (LDC). • The phonetic and prosodic marking of turns has a limited impact on minimizing disaffiliation (PDC).

17 From well-managed or no display of disaffiliation to bald-on-record display of disaffiliation.

Band 2 Concerning	• A disaffiliative stance can be imputed to the speaker and interactional continuity can be disrupted. The use of disaffiliative actions and underuse of strategies can create or escalate tension (DAA). • There is recruitment of disaffiliative linguistic devices while the ones constituting a dispreferred format are inadequately represented (LDC). • The speaker's talk contains disaffiliative segmental and suprasegmental features (PDC).
Band 1 Intervention needed	• Social solidarity is jeopardized due to the deployment of ostensible disaffiliative actions. The speaker's approach projects a clearly disaffiliative and uncooperative stance (DAA). • The design of turns lacks basic dis-preferenced features. Highly disaffiliative linguistic devices exist (LDC). • Turns are marked by pronounced disaffiliative segmental and suprasegmental features (PDC).

Affiliation promotion[18]

Band 5 Exemplary	• The speaker is maximally pro-social through successful launches of substantively affiliative actions (AFA). • An impressive array of lexical, morphosyntactic, phonetic and prosodic elements are skillfully mobilized in designing preferred actions (ATD). • A very high degree of empathetic understanding and sharing of frames is evoked, established, and maintained via the use of substantive forms of empathy display and frame identification (EAF). • The speaker is extremely keen and competent in pursuing, maintaining, and restoring intersubjectivity (ITS).

18 From overt display of affiliation to no show of affiliation.

Band 4 Good	• Social solidarity is supported through the employment of commonly used affiliative actions (AFA). • The speaker can use (non)lexical devices to display affiliation. Phonetic and prosodic marking are in general in tune with turns (ATD). • There is sufficient recognition, validation and understanding of the interlocutor's emotions and frames (EAF). • Threats to mutual understanding are adequately addressed in the interaction (ITS).
Band 3 Average	• There is some use of affiliative actions to support the interlocutor's affective stance (AFA). • Turn design displays some conventionalized (non)verbal affiliation promotion (ATD). • Attempts are made to affiliate with the interlocutor's emotions and frames, though not always successful (EAF). • Intersubjectivity is sometimes checked and defended (ITS).
Band 2 Concerning	• Affiliative actions are underused. The interlocutor can feel somewhat unendorsed or unsupported (AFA). • (Non)verbal affiliative devices can be inadequate or mis-designed (ATD). • Little display of empathetic understanding of the interlocutor's feelings and experiences (EAF). • The speaker's understanding is prioritized while common understanding is largely unaddressed (ITS).
Band 1 Intervention needed	• Little to no indexation of affiliation. The interlocutor's affective needs are disregarded (AFA). • Preferenced (non)lexical affiliative features[19] are rare to non-existent. Little to no prosodic matching or upgrading for affiliative work (ATD). • Noticeable misalignment from and disaffiliation with the interlocutor's emotions or frames (EAF). • The speaker overtly privileges their understanding and is disinterested in achieving intersubjectivity (ITS).

19 Preferred, preferenced and affiliative features are supportive of social solidarity in both 1st pp or 2nd pp positions.

Morality

Band 5 Exemplary	• The speaker's manner of interaction is maximally cooperative, fully sustaining the endemic moral order of interaction (EMO). • The interaction clearly indexes universally preferred moral qualities in the speaker (UMO). • The speaker's moral membership is fully established through the demonstration of moral conduct specific to the standards in their community (CMO).
Band 4 Good	• The moral order of interaction is well sustained through a focus on cooperation instead of constraint (EMO). • The speaker's conduct can allow for the ascription of some universally preferred moral qualities in the speaker (UMO). • The speaker's conduct showcases a good understanding of moral standards known to members of the shared community (CMO).
Band 3 Average	• Interactional order is in general maintained though some conducts are not fully morally accountable or accounted for (EMO). • There is no strong indexation of either universally preferred or dispreferred moral qualities (UMO). • Community-specific moral conduct is lacking, though the speaker's moral membership is still accepted (CMO).
Band 2 Concerning	• Interactional order is threatened by noticeable moral breaches (EMO). • Some universally dispreferred moral qualities can be ascribed to speaker (UMO). • The speaker's conduct is recognizably morally questionable to members in the community (CMO).
Band 1 Intervention needed	• Moral order is disrupted by a constraining interactional manner and morally unaccountable conduct (EMO). • The speaker deploys universally morally sanctionable actions that challenge their status as a moral being (UMO). • The speaker faces ostracization due to the display of morally unjustifiable conduct in the community (CMO).

Reasoning

Band 5 Exemplary	• Extremely efficacious, multi-angled and context-fitting remedies are provided to address interactional troubles at both the action and project levels (IAR). • The speaker provides exceedingly well-reasoned and contextually defensible accounts at both the action and project levels (IAA). • The entire occasion of interaction is exceptionally well organized at both the sequential and sequence levels, providing maximal progressivity and projectability (OSO).
Band 4 Good	• The speaker proffers quality and applicable remedies to troubles, though slightly more substantiation of the remedies would be ideal (IAR). • The accounts proffered conform with the shared reasoning in a normative social life (IAA). • Interaction is structured in a recognizable manner around a beginning, different units of interaction and a closing (OSO).
Band 3 Average	• The remedies offered have limited impact on addressing the troubles in the interactional context (IAR). • Accounts are provided but are not the most readily acceptable ones or are in need of development (IAA). • The overall structural organization is in place, though the design of some interactional units can be improved (OSO).
Band 2 Concerning	• The remedies are simplistic or ineffective. Interactional troubles are largely unaddressed (IAR). • Noticeable limitations exist in the accounts supplied, dis-conforming with the shared reasoning among interactants (IAA). • There are noticeable issues with the organization of sequences and the organization of sequences of sequences (OSO).
Band 1 Intervention needed	• Either no remedies are provided, or the remedies provided are problematic or context-misfitting (IAR). • Either no accounts are proffered, or the accounts supplied are poorly reasoned or norm-defying (IAA). • The organization of interaction is weak, actions are unrecognizable, and progressivity is stalled (OSO).

Social role management

Band 5 Exemplary	• The speaker has outstanding competence in enacting and orienting to social roles that are highly congruent with their and the interlocutors' categories, matching their conduct with their respective category-bound predicates and relative hierarchical positioning (CAP). • There is an excellent application of the standardized relational pairs, membership categorization devices and duplicative organizations to which the speaker and interlocutor belong (BMA). • Highly skilled mediation and prioritization of the speaker's and the interlocutors' roles are demonstrated (SRC).
Band 4 Good	• The speaker demonstrates the ability to enact and orient to relevant social roles. The matching of category-bound predicates is overall felicitous (CAP). • The speakers can utilize some broader membership apparatuses (BMA) to make their conduct recognizable (BMA). • Role competition is balanced to achieve successful interaction (SRC).
Band 3 Average	• The expected roles are in general enacted and oriented to. Some predicates can be over-realized or under-realized (CAP). • The application of BMA is limited but no misuse exists (BMA). • The speaker's and the interlocutors' primary roles are oriented to, but other roles are insufficiently addressed (SRC).
Band 2 Concerning	• Social role management is insufficient. Categories are not well matched with predicates (CAP). • There are incidents of misapplications of BMA, suggesting a lack of knowledge of what context-fitting BMA to draw on (BMA). • Primary roles are not adequately attended to (SRC).

Band 1 Intervention needed	• Normatively expectable categories are not attended to and there is grave misunderstanding regarding category-bound predicates (CAP). • The speaker neglects context-relevant BMA or seriously mismanaged BMA, disrupting the interaction (BMA). • There is mis-prioritization of roles and role competition is overlooked (SRC).

Chinese version

一：控制冲突

五分：闪光	• 考生能够非常成功地巧妙地化解伤面子的情况，人际关系[20]完全没有损伤。可能伤面子的话要么没说，要么说得非常婉转，完全不会冒犯到别人。 • 考生说话时有技巧地使用了丰富的婉转修饰不确定性的词语和句型结构，非常有效地减缓冲突。 • 考生在控制冲突时能非常好地通过语音、语调和语气来减缓伤面子的行为。
四分：良好	• 考生能较好地控制伤面子的情况。非常伤面子的行为能委婉处理，有些不是太伤面子的行为可能处理得较为直接，但在特定语境下是可以接受的。 • 考生能很好地使用实词和虚词来化解冲突。词不达意的情况很少，有的话也不会影响交流。 • 考生控制冲突时的语音、语调、语气和说话内容相匹配，能起到舒缓冲突的作用。
三分：一般	• 考生在控制冲突时有时说话比较直接，有时又比较婉转，不太统一，但总体上能让人接受，不会伤到别人或损害人际关系。 • 考生能使用一些模式化的婉转语言来缓解冲突至一个可以接受的范围。 • 考生控制冲突时的语音、语调、语气对控制冲突有比较有限的帮助。

20 人际关系，别人不一定特指交流的另一方，可能是其他相关的人员和广义的社会关系。

二分：有缺陷	• 考生有的话说得比较直接，会给交流造成阻碍。考生有时交流技巧不足，容易制造和激化矛盾。 • 考生婉转的词汇和句型使用得不够，有的语言会伤害到别人的面子。 • 考生在控制冲突时，语音语调语气里有时会让对方觉得有点伤人。
一分：需努力	• 考生使用了伤人的语言，给人际关系造成了损害。 • 考生有的语言缺少基本的婉转修饰词语和句型结构，有的话会得罪人。 • 考生在应该控制冲突时，语音语调语气里却出现了会冲突冒犯到别人的成分。

二：拉近关系

五分：闪光	• 考生能通过一系列的语言和行为来非常成功有效地拉近和对方的情感距离，增进人际关系。 • 考生在拉近情感时在词法、句法、语音、语调、语气上都处理得很有技巧。 • 考生在拉近关系时展示出很强的共情同理心，能换位思考，充分体谅到别人的情感和处境。 • 考生在拉近关系时展现出非常强的沟通、征询、商量和共鸣[21]的意识。
四分：良好	• 考生能有效使用常用的拉近情感的言行来维护人与人之间的情面。 • 考生能使用实词和虚词来拉近情感，拉近情感时语音语调与措辞相匹配。 • 考生在拉近关系时能有效地留意、认可、理解对方的情绪和角度。 • 考生在拉近关系时展现出良好的沟通能力来和对方取得共鸣。

21 共鸣不一定是要完全一致，而是能理解对方的想法和角度。

三分： 一般	• 考生有一些拉近关系的行为，虽然比较有限但还是照顾到了对方的情感需求。 • 考生组织语言时能使用一些模式化的实词，虚词和语气特征来拉近和对方的情感距离。 • 考生在拉近情感时有去体谅对方的情绪和处境，虽然并非完全到位，但总体尚可。 • 考生拉近情感时有时会努力去理解对方，和对方取得共鸣。
二分： 有缺陷	• 考生拉近情感的言行相对而言比较少，对方会觉得考生对自己不够上心和支持。 • 考生措辞和语气上拉近情感的效果不足，或者有用得不合适的地方。 • 在适合拉近情感的时刻，考生没有足够地体谅对方的情绪和角度。 • 考生有时优先考虑自己对事物的认识，而不是去通过努力和对方获得共鸣来拉近情感。
一分： 需努力	• 考生几乎没有拉近情感的言行，很大程度上忽视了对方的情感需求。 • 考生的语言无论在用词、句型、或者语气语调上几乎没有什么拉近情感的成分。 • 考生在应该拉近情感的时刻忽视了对方的情感和视角。 • 考生在可以拉近关系的时刻专注于自己对问题的认识，没有有效地和对方取得共鸣。

三：人品素质[22]

五分： 闪光	• 考生在交流中很明显地展现出普世的优秀道德品质。 • 考生的交流强烈体现出中文文化和语境下值得称赞的美德。 • 考生在交流过程中非常真诚地对话，非常有诚意地推动对话进展，体现出对交流方极大的尊重。
四分： 良好	• 考生的言行展现出良好的个人素质。 • 考生的表现和中文道德素质观里的一些优良品质相对应。 • 整个交谈中考生尊重对方，态度端正，推进交流。

22 拉近情感更多的是外在表现，道德人品更多是内在个人素质

三分： 一般	• 考生在交流中体现了常识性的，符合社会常规的人品素质，没什么特别值得称赞的地方，也没有什么值得质疑的地方。 • 考生的言行大体符合中文文化的道德观，没有什么特别的优秀品质，但也不会引人非议。 • 考生交流时大体上态度端正，虽然有时主观能动性有些欠缺，但能保证交流可以顺利进行。
二分： 有缺陷	• 考生的某些言行会让大众对考生的为人和品德产生疑问。 • 中文语境下考生的某些言行会让别人对考生的个人素质提出一些质疑。 • 考生在交流中有出现态度不够端正，不够尊重对方的地方，会影响到交流顺畅进行。
一分： 需努力	• 考生的一些表现会让人负面评价他的素质和人品。 • 考生的言行有与中国文化语境中道德观和为人品质相冲突的地方。 • 考生交流中有不尊重对方的地方，导致交流无法深入进行。

四：理性思辨

五分： 闪光	• 考生不论是在交流过程中还是在交流目的上都能多角度地提供非常有效，符合语境的解决问题的方法。 • 不论是在交流过程中还是在交流目的上考生都能为自己的行为和论点提供非常合理而充分的理由。 • 考生的交流具有起承转合，不论是在句子还是整体层面都结构妥帖，层层深入，易于理解。
四分： 良好	• 考生能提供有效的可行的解决问题的方法，虽然方法再多展开一些方可论完美。 • 考生解释自己行为和观点的理由符合大众常规的理性思维。 • 考生的交流有一定的结构，有开头，细节上有展开，有收尾。

三分： 一般	• 考生提供了解决问题的方法，虽然不一定是最优解，但仍然可以解决问题。 • 考生为自己行为和观点提供了理由，虽然理由不一定是最容易令人接受的理由，或者理由需要更多补充，但理由总体还是可以接受的。 • 考生的交流有大体的结构框架，虽然交流细节上还可以改进。
二分： 有缺陷	• 在遇到交流困难时，考生提供的解决问题的方法有些简单或者效果不能算好。交流障碍没有得到很好的处理。 • 考生为自己的行为和观点找的理由存在一定的逻辑漏洞，不太符合大众的理性推断。 • 考生的交流在句子和整体层面上存在一些结构性问题。
一分： 需努力	• 考生或是没有提供解决问题的方法，或者是提供的方法不合适。 • 考生或是没有为自己的行为和观点提供理由，或者是提供的理由不合常理。 • 考生的交流整体结构存在不合理因素，行为目的不明确，交流难以顺畅进行。

五：身份意识

五分： 闪光	• 考生对考题中自己的身份和对方的身份把握得非常准确，理解双方身份带来的权利，责任和义务。考生的言行与自己的身份和对方的身份相当一致，非常符合两人之间的辈分等级关系。 • 考生能非常有效地利用两人之间的社会关系以及和其他人之间的社会群体关系来有效地进行交流。 • 考生能非常娴熟地处理和平衡自己和对方的多重身份。
四分： 良好	• 考生能处理自己和对方的身份所带来的约束，考生能做到根据自己和对方的身份来调整自己的言行。 • 考生能通过广义社会群体关系来推进交流。 • 考生能有效处理身份之间的冲突来取得顺利交流。
三分： 一般	• 考生对考题中自己和别人的身份有一定认知。考虑到双方的身份，考生的言行有时会有些过了或者不足的地方，但总体还算符合身份。 • 考生对广义社会群体关系使用有限，但无明显误用的地方。 • 考生对自己和对方的主体身份有一定认知，但对其他的身份可能处理得不够到位。

二分：有缺陷	• 考生在对自己和别人的身份处理上有不足之处，考生的言行与双方的身份，辈分，等级不够协调。 • 考生有误用广义社会群体关系的地方，体现出他对大环境下人与人之间的关系缺乏应有的理解。 • 考生自己和对方的主体身份有处理得不到位的地方。
一分：需努力	• 考生对自己和对方最基本的身份和相对应的言行存在误解。 • 考生或者忽视了双方所处在的广义社会群体关系，或者是完全误用了广义的关系，阻碍了交流。 • 考生错误地优先处理了次要的身份，忽视了平衡身份的重要性。

Appendix V: The self-assessment questionnaire

English version

Section A: Test-taking experience

1. The tasks in this test can help me develop stronger interpersonal communication skills if I practise them often.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

2. In real life I would behave differently from how I did in the test.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

3. I think other types of Chinese speaking tests are much better than the one I just tried.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

4. This test I just had says a lot about my ability to handle difficult interpersonal situations in Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

5. My performances on the test are similar to how I would behave in similar situations in real life.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

6. Organisations (e.g., universities and workplaces) should use the speaking test I just tried to assess second language Chinese speakers' ability to interact in Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

7. The speaking test I just had should become more mainstream on the market.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

8. The results from this test can inform others of my ability to address difficult interpersonal situations in Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

9. The test offers a poor prediction of a person's ability to tackle difficult interactions in Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

10. Chinese language teachers should introduce similar tasks like the ones I just tried into their classes.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

11. If someone scores high on this test, I would say they can communicate appropriately in Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

Section B: Chinese interaction information

This section asks you to give your assessment of your ability to interact appropriately in spoken Chinese with speakers of Chinese. Please respond to the questions by choosing the option that you think most accurately describes your ability in specific settings.

It might be possible that you are unfamiliar with some of the situations in the questions, but you are encouraged to assess your hypothetical performance based on your knowledge of how you interact in other more familiar situations and the feedback you have received from your Chinese speaking friends. If you are still unable to assess your ability, please select "unable to assess".

1. I can adjust my spoken Chinese to the social role of my interlocutor (e.g., I know how to talk appropriately to a Chinese professor at university).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

2. I know how to choose Chinese words carefully to avoid offending speakers of Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

3. I can communicate in spoken Chinese appropriately based on my social role in specific settings (e.g., talking appropriately as a student in a school setting).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

4. When I tell a story in Chinese, I can follow a clear structure (e.g., beginning, development, narrowing-down and conclusion).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

5. When talking in Chinese, I use language to show agreement so as to build rapport (e.g., "en", "yes" and "I agree"/例如"嗯""对的""我很同意"等).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

6. When talking in Chinese, my intonation is appropriate according to the interactional settings.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

7. I think my spoken Chinese sometimes lacks enough empathy and sympathy as required in the Chinese culture. (e.g., not showing enough that I care about the other person)

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

8. I noticed my spoken Chinese sometimes sounds overly direct and forceful.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

9. I can change my spoken Chinese effectively when my social role changes (e.g., talking as a student to my Chinese teacher vs talking as a friend to my Chinese friend).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

10. When I speak Chinese with a few Chinese speakers in a group setting, I always come across as a collaborative team player.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

11. I notice my spoken Chinese sometimes appears too casual and not serious enough in the Chinese context.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

12. When communicating in spoken Chinese, I always look after my interlocutor's emotions and their ability to handle my message, especially if what I am saying might be distressing to them.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

13. My spoken Chinese can convey my respect for my interlocutor.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

14. When I need to explain reasons for my actions in spoken Chinese, sometimes I feel that my explanation is not detailed enough.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

15. I can communicate in spoken Chinese in a calm and collected manner when faced with situations that can offend my interlocutor.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

16. When communicating in Chinese, I communicate in a manner that shows I affirm and support my interlocutor.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

17. When I experience negative emotions (e.g., angry or upset) during my communication in spoken Chinese, I can control my emotions well and talk calmly.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

18. I can adapt my spoken Chinese effectively when I am talking to speakers who have different social roles (talking to a Chinese neighbour vs talking to my Chinese boss at work).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

19. I know how to use language that sounds sincere to de-escalate when I sense I can potentially offend speakers of Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

20. I can elaborate my ideas and opinions comprehensively in spoken Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

21. I know how to use indirect language in spoken Chinese to manage difficult interpersonal situations.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

22. I notice sometimes I reject/disagree with Chinese speakers' opinions in a manner that is considered impolite by them.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

23. When I need to provide solutions to problems, I can propose reasonable and sensible solutions in spoken Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

24. When faced with difficult situations, I am very solution-focused in my spoken Chinese.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

25. When speaking Chinese, I focus on negotiating and reaching an agreement with my interlocutor.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

26. When I speak Chinese, I have a good understanding of the obligations, rights and responsibilities in my interlocutor's social role (e.g., If I am an employee in a Chinese company, I know what my Chinese manager can expect from me).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

27. When expected to provide reasons for my action, I can provide believable and appropriate reasons in spoken Chinese (e.g., why I can't go to my friend's birthday party)

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

28. When speaking in Chinese, sometimes I notice that I don't have a deep understanding of the seniority and social ranking systems in Chinese society.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

29. I can communicate in a caring and concerned manner in Chinese so as to maintain strong interpersonal relationships.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

30. When encountering difficulties, I can propose solutions in Chinese that sound genuine and sincere.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

31. When communicating in Chinese, I never use language that can sound overly direct or forceful.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

32. I notice that when communicating in Chinese, sometimes I appear rather passive when there is a problem (e.g., not being proactive enough in problem-solving).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

33. When I talk in Chinese, I think I come across as an honest person to my interlocutor.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

34. When speaking in Chinese, I have a good understanding of the obligations, rights and responsibilities in my social role (e.g., If I am talking as a student, I know what I should and should not do in the Chinese context).

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

35. Other Chinese speakers say my spoken Chinese sometimes sound threatening to them.

Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree	Unable to assess

Chinese version

问卷调查

A: 测试体验

1. 如果经常练习的话，今天考试里的题目能很好地帮助我提高我的中文人际交流能力。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

2. 在现实生活中如果遇到了和考试里相似的场景，我的交流方式会和我在考试时的方式有很大的差别。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

3. 和今天的汉语口语考试相比，我更喜欢其他类型的口语考试。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

4. 今天的考试能很好地测量我用中文处理复杂人际关系的能力。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

5. 在现实生活中如果遇到了和考试里相似的场景，我的表现会和我在考试里的表现很相似。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

6. 机构（例如大学或者公司）应该多使用我刚才参加的测试来衡量一个人的中文交流能力。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

7. 我希望我刚才参加的汉语口语测试能在市场上更常见。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

8. 我觉得别人能从我的考试成绩中了解我用中文处理复杂人际关系的能力。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

9. 这个测试无法推测一个人用中文处理复杂交流问题的能力。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

10. 中文老师应该多在课堂上使用我刚才在测试中尝试的练习。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

11. 如果一个人在我刚才做的测试里得了高分，我觉得那个人能很得体地用中文交流。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意

B:中文表达能力

接下来的问题请你评价你用中文和别的中文使用者交流的能力。请在每一个问题中选择你认为最能代表你的能力的选项。有时你可能不熟悉你在某个问题中的能力和表现，在这种情况下，请你根据自己平时的汉语交流能力的了解进行类推。如果你因为不了解相关背景而实在无法对某一题里的表现作出评价的话，请选择“无法评价”。

1. 我能根据说话的对象选择合适得体的语言（例如：如何得体地和大学老师交流）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

2. 在可能冒犯到别人的时候，我能够小心地选择中文词汇来避免这种情况。

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

3. 我能根据自己的身份得体地与他人用中文交流（例如：作为一个学生在校园里应该怎么说话）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

4. 我用中文叙述事情时很有条理（例如一个故事有开始，发展，收尾，结束）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

5. 当我用中文交流时，我会经常赞同别人的观点（例如说“嗯”“对的”“我很同意”等）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

6. 当我说中文时，我的语气总是很得体

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

7. 当我说中文时，我感觉我有时缺乏对他人的共情同理心（例如我没有去站在对方的角度考虑问题）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

8. 我觉得当我用中文交流时，我的语言有时会过于直接强硬

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

9. 我能很好地根据自己身份的变化来调整自己的语言（例如：作为一个学生该怎么和老师说话，以及作为一个朋友该怎么和自己的朋友说话）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

10. 当我在一群说中文的人里说中文时，我能表现出自己是一个很有团队合作意识的人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

11. 当我和朋友用中文交流时，我觉得自己的态度有时会比较随便，不认真

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

12. 当我用中文交流时，我会照顾对方的情绪和承受能力（例如使用安抚他人的语言）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

13. 当我用中文交流时，我会表现出我很尊重和我说话的人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

14. 当我用中文解释一件事时，有时我会觉得我的解释内容不够具体

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

15. 当遇到有可能会伤和气的事情，我能用中文平和地与别人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

16. 当我说中文时我会经常肯定鼓励别人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

17. 当我用中文交流时，我能很好地控制自己的情绪，冷静地与他人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

18. 我知道如何用中文对什么样的人说什么样的话，能有效地根据说话的对象调整自己的语言（例如：和邻居说话和对上司说话之间的区别）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

19. 当我觉得我可能会得罪别人时，我能用真诚的语言来化解矛盾

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

20. 当我用中文交流时，我能够充分完整地表达自己的观点和意见

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

21. 为了化解矛盾，我能用中文婉转地与他人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

22. 当我用中文交流时，有时我会以不礼貌的方式反对别人的观点

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

23. 当需要提供解决问题的方法时，我能用中文提出很合适得体的方法

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

24. 当遇到困难时，我能用中文表现出我以解决问题为重的意图

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

25. 当我说中文时，我能展现出很强的沟通的意识（例如采取询问和商量的口吻）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

26. 当我说中文时，我很了解中国文化下每个人不同的责任、权利和义务（例如：在中国公司里上司和下属间不同的责任、权利和义务）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

27. 当有需要时，我能用中文为自己的行为提供合理的解释（例如：为什么不能赴约）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

28. 当我用中文交流时，有时我觉得我对中文语境下人与人之间的辈分和等级观念缺乏深入的了解

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

29. 我知道如何用中文来关心叮嘱别人，达到维持良好的人际关系的目的

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

30. 在遇到困难时，我能用中文提出很有诚意的解决问题的方法

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

31. 当我用中文交流时，我不会使用过于直接强硬的语言

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

32. 当我用中文交流时，有时我觉得我处理问题的态度有些消极（例如不去主动地想办法解决问题）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

33. 当我说中文时，我觉得我显示出自己是一个坦诚的人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

34. 当我说中文时，我对自己身份的定位很清楚，了解自己的身份所带来的责任、权利和义务（例如：在中国语境下一个学生该做什么和不该做什么）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

35. 当我用中文交流时，有时我会使用威胁性的语言

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法评价

Appendix VI: The peer-assessment questionnaire

汉语交际能力问卷调查
您好！这份问卷调查请您评价您的外国朋友用中文和别的中文使用者交流的能力。请在每一个问题中选择您认为最能代表您的外国朋友能力的选项，您可以选择加粗您的答案，例如：

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

有时您可能不熟悉您的朋友在某个问题中的能力和表现，在这种情况下，请您根据平时对您朋友的汉语交流能力的了解进行类推。如果您因为不了解相关背景而实在无法对您的朋友在某一题里的表现作出评价的话，请选择"无法评价"。您的回答会有助于我们设计更好的汉语教学和测试材料，非常感谢您的协助。
您的外国朋友的名字_____

1. 我的外国朋友能根据说话的对象选择合适得体的语言（例如：如何得体地和大学老师交流）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

2. 我的外国朋友为了避免冒犯到说中文的人，知道如何斟酌用词

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

3. 我的外国朋友能根据自己的身份得体地与他人用中文交流（例如：作为一个学生在校园里应该怎么说话）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

4. 我的外国朋友用中文叙述事情时很有条理（例如一个故事有开始，发展，收尾，结束）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

5. 当我的外国朋友用中文交流时，他/她会经常赞同别人的观点（例如说“嗯”“对的”“我很同意”等）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

6. 当我的外国朋友说中文时，他/她的语气总是让我觉得很得体

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

7. 我的外国朋友说中文时，我感觉他/她有时缺乏对他人的共情同理心（例如他/她不去站在对方的角度考虑问题）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

8. 我的外国朋友用中文交流时，措辞有时会过于直接强硬

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

9. 我的外国朋友能很好地根据自己身份的变化来调整自己的语言（例如：作为一个学生该怎么和老师说话，以及作为一个朋友该怎么和自己的朋友说话）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

10. 当我的外国朋友在一群说中文的人里说中文时，我能感受到他/她是一个很有团队合作意识的人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

11. 当我和我的外国朋友用中文交流时，我有时觉得他/她的态度会比较敷衍随便

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

12. 当我的外国朋友用中文交流时，他/她能照顾对方的情绪和承受能力（例如适时使用安抚他人的语言）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

13. 当我和我的外国朋友用中文交流时，我觉得他/她很尊重我

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

14. 当我的外国朋友用中文解释事情的缘由时，有时我会觉得他/她的解释内容很单薄

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

15. 当遇到有可能会伤和气的事情，我的外国朋友会用中文平和地与别人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

16. 当我的外国朋友说中文时他/她会经常肯定鼓励别人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

17. 我的外国朋友用中文交流时，能很好地控制自己的情绪，冷静地与他人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

18. 我的外国朋友知道如何用中文对什么样的人说什么样的话，能有效地根据说话的对象调整自己的语言（例如：和邻居说话和对上司说话之间的区别）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

19. 当我的外国朋友觉得可能会得罪别人时，他/她会用真诚的语言来化解矛盾

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

20. 当我的外国朋友用中文交流时，他/她能够充分完整地表达自己的观点和意见

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

21. 我的外国朋友为了化解矛盾，能用中文婉转地与他人交流

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

22. 我的外国朋友用中文交流时，有时会以不礼貌的方式反驳别人的观点

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

23. 当需要提供解决问题的方法时，我的外国朋友用中文提出的方法让我觉得很中肯

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

24. 当遇到困难时，我的外国朋友能用中文展现出他/她以解决问题为重的意图

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

25. 当我的外国朋友说中文时，他/她展现出很强的沟通的意识（例如采取征询和商量的口吻）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

26. 当我的外国朋友说中文时，我觉得他/她理解中国文化下每个人不同的责任、权利和义务（例如：在中国公司里上司和下属间不同的责任、权利和义务）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

27. 当有需要时，我的外国朋友能用中文为自己的行为提供合理的解释（例如：为什么不能赴约）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

28. 我的外国朋友用中文交流时，有时我觉得他/她对中文语境下人与人之间的辈分和等级观念缺乏深入的了解

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

29. 我的外国朋友知道如何用中文来关心叮嘱别人，达到维持良好的人际关系的目的

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

30. 在遇到困难时，我的外国朋友能用中文提出很有诚意的解决问题的方法

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

31. 我的外国朋友用中文交流时，不会使用过于直接强硬的语言

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

32. 我在和我的外国朋友用中文交流中，有时我会觉得他/她处理问题的态度有些消极（例如不去主动地想办法解决问题）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

33. 当我的外国朋友说中文时，我觉得他/她是一个坦诚的人

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

34. 当我的外国朋友说中文时，我觉得他/她对自己身份的定位很清楚，了解自己的身份所带来的责任、权利和义务（例如：在中国语境下一个学生该做什么和不该做什么）

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

35. 我的外国朋友用中文交流时，有时会使用威胁性的语言

非常不同意	不同意	有点不同意	有点同意	同意	非常同意	无法 评价

Author Information

Dr. David Wei Dai
ORCiD: https://orcid.org/0000-0002-3575-131X
Institutional address: Room W8.16, 20 Bedford Way, UCL Institute of Education, University College London, London, WC1H 0AL, UK
Email: david.dai@ucl.ac.uk
Twitter/X: @drdavidweidai

Language Testing and Evaluation

Series editors: Claudia Harsch and Günther Sigott

Vol. 1 Günther Sigott: Towards Identifying the C-Test Construct. 2004.

Vol. 2 Carsten Röver. Testing ESL Pragmatics. Development and Validation of a Web-Based Assessment Battery. 2005.

Vol. 3 Tom Lumley: Assessing Second Language Writing. The Rater's Perspective. 2005.

Vol. 4 Annie Brown: Interviewer Variability in Oral Proficiency Interviews. 2005.

Vol. 5 Jianda Liu: Measuring Interlanguage Pragmatic Knowledge of EFL Learners. 2006.

Vol. 6 Rüdiger Grotjahn (Hrsg. / ed.): Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, Empirical Research, Applications. 2006.

Vol. 7 Vivien Berry: Personality Differences and Oral Test Performance. 2007.

Vol. 8 John O'Dwyer: Formative Evaluation for Organisational Learning. A Case Study of the Management of a Process of Curriculum Development. 2008.

Vol. 9 Aek Phakiti: Strategic Competence and EFL Reading Test Performance. A Structural Equation Modeling Approach. 2007.

Vol. 10 Gábor Szabó: Applying Item Response Theory in Language Test Item Bank Building. 2008.

Vol. 11 John M. Norris: Validity Evaluation in Language Assessment. 2008.

Vol. 12 Barry O'Sullivan: Modelling Performance in Tests of Spoken Language. 2008.

Vol. 13 Annie Brown / Kathryn Hill (eds.): Tasks and Criteria in Performance Assessment. Proceedings of the 28th Language Testing Research Colloquium. 2009.

Vol. 14 Ildikó Csépes: Measuring Oral Proficiency through Paired-Task Performance. 2009.

Vol. 15 Dina Tsagari: The Complexity of Test Washback. An Empirical Study. 2009.

Vol. 16 Spiros Papageorgiou: Setting Performance Standards in Europe. The Judges' Contribution to Relating Language Examinations to the Common European Framework of Reference. 2009.

Vol. 17 Ute Knoch: Diagnostic Writing Assessment. The Development and Validation of a Rating Scale. 2009.

Vol. 18 Rüdiger Grotjahn (Hrsg. / ed.): Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research. 2010.

Vol. 19 Fred Dervin / Eija Suomela-Salmi (eds. / éds): New Approaches to Assessing Language and (Inter-)Cultural Competences in Higher Education / Nouvelles approches de l'évaluation des compétences langagières et (inter-)culturelles dans l'enseignement supérieur. 2010.

Vol. 20 Ana Maria Ducasse: Interaction in Paired Oral Proficiency Assessment in Spanish. Rater and Candidate Input into Evidence Based Scale Development and Construct Definition. 2010.

Vol. 21 Luke Harding: Accent and Listening Assessment. A Validation Study of the Use of Speakers with L2 Accents on an Academic English Listening Test. 2011.

Vol. 22 Thomas Eckes: Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments. 2011. 2nd Revised and Updated Edition. 2015.

Vol. 23 Gabriele Kecker: Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen. 2011.

Vol. 24 Lyn May: Interaction in a Paired Speaking Test. The Rater´s Perspective. 2011.

Vol. 25 Dina Tsagari / Ildikó Csépes (eds.): Classroom-Based Language Assessment. 2011.

Vol. 26 Dina Tsagari / Ildikó Csépes (eds.): Collaboration in Language Testing and Assessment. 2012.

Vol. 27 Kathryn Hill: Classroom-Based Assessment in the School Foreign Language Classroom. 2012.

Vol. 28 Dina Tsagari / Salomi Papadima-Sophocleous / Sophie Ioannou-Georgiou (eds.): International Experiences in Language Testing and Assessment. Selected Papers in Memory of Pavlos Pavlou. 2013.

Vol. 29 Dina Tsagari / Roelof van Deemter (eds.): Assessment Issues in Language Translation and Interpreting. 2013.

Vol. 30 Fumiyo Nakatsuhara: The Co-construction of Conversation in Group Oral Tests. 2013.

Vol. 31 Veronika Timpe: Assessing Intercultural Language Learning. The Dependence of Receptive Sociopragmatic Competence and Discourse Competence on Learning Opportunities and Input. 2013.

Vol. 32 Florian Kağan Meyer: Language Proficiency Testing for Chinese as a Foreign Language. An Argument-Based Approach for Validating the Hanyu Shuiping Kaoshi (HSK). 2014.

Vol. 33 Katrin Wisniewski: Die Validität der Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen. Eine empirische Untersuchung der Flüssigkeitsund Wortschatzskalen des GeRS am Beispiel des Italienischen und des Deutschen. 2014.

Vol. 34 Rüdiger Grotjahn (Hrsg./ed.): Der C-Test: Aktuelle Tendenzen/The C-Test: Current Trends. 2014.

Vol. 35 Carsten Roever / Catriona Fraser / Catherine Elder: Testing ESL Sociopragmatics. Development and Validation of a Web-based Test Battery. 2014.

Vol. 36 Trisevgeni Liontou: Computational Text Analysis and Reading Comprehension Exam Complexity. Towards Automatic Text Classification. 2015.

Vol. 37 Armin Berger: Validating Analytic Rating Scales. A Multi-Method Approach to Scaling Descriptors for Assessing Academic Speaking. 2015.

Vol. 38 Anastasia Drackert: Validating Language Proficiency Assessments in Second Language Acquisition Research. Applying an Argument-Based Approach. 2015.

Vol. 39 John Norris: Developing C-tests for estimating proficiency in foreign language research. 2018.

Vol. 40 Günther Sigott (Ed./Hrsg): Language Testing in Austria: Taking Stock/Sprachtesten in Österreich: Eine Bestandsaufnahme. 2018.

Vol. 41 Carsten Roever / Gillian Wigglesworth (Eds.): Social perspectives on language testing. Papers in honour of Tim McNamara. 2019.

Vol. 42 Klaus Siller: Predicting Item Difficulty in a Reading Test. A Construct Identification Study of the Austrian 2009 Baseline English Reading Test 2020.

Vol. 43 Anastasia Drackert / Mirka Mainzer-Murrenhoff / Anna Soltyska / Anna Timukova (Hrsg.): Testen bildungssprachlicher Kompetenzen und akademischer Sprachkompetenzen. Zugänge für Schule und Hochschule. 2020.

Vol. 44 Khaled Barkaoui: Evaluating Tests of Second Language Development. A Framework and an Empirical Study. 2021.

Vol. 45 Theresa Weiler: Testing Lexicogrammar. An Investigation into the Construct Tested in the Language in Use Section of the Austrian Matura in English. 2022.

Vol. 46 Charalambos Kollias: Virtual Standard Setting: Setting Cut Scores. 2023.

Vol. 47 Nikola Dobrić / Hermann Cesnik / Claudia Harsch (eds.): Festschrift in Honour of Günther Sigott: Advanced Methods in Language Testing. 2023.

Vol. 48 David Wei Dai: Assessing Interactional Competence. Principles, Test Development and Validation through an L2 Chinese IC Test. 2024.

www.peterlang.com