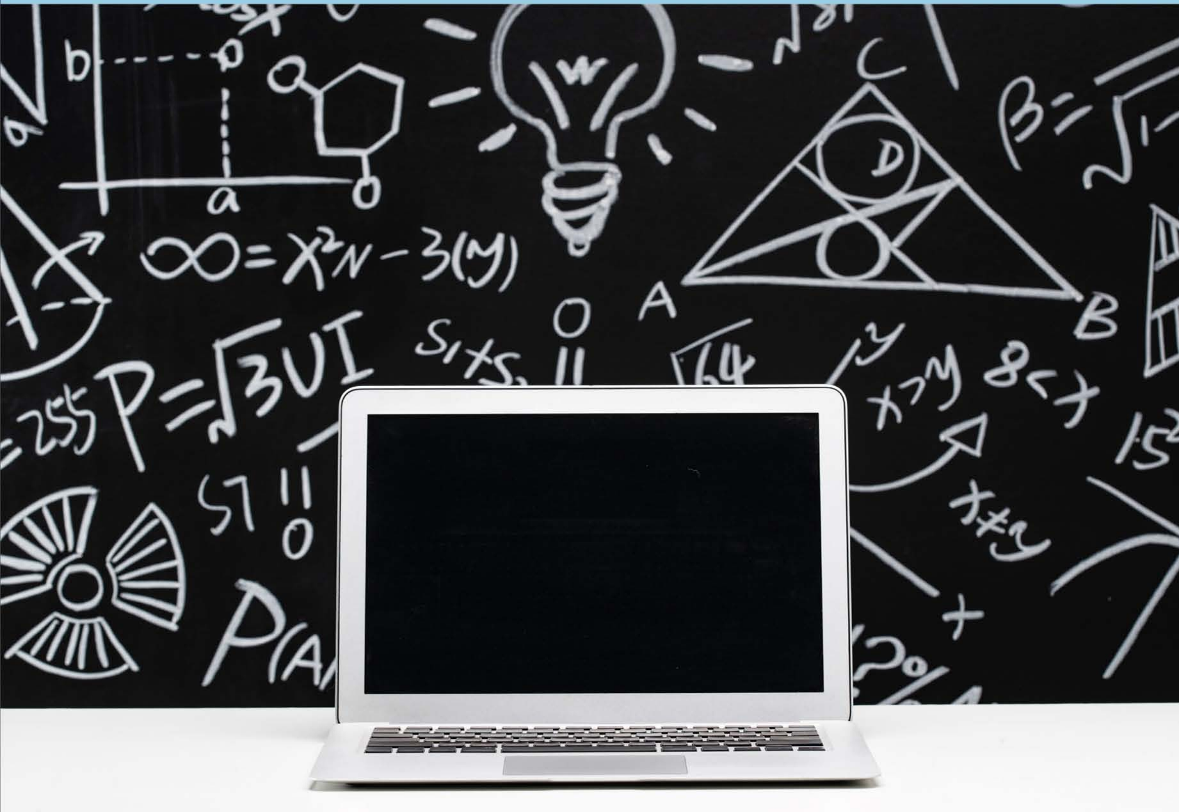
Edited by Eirini Geraniou, Cosette Crisan, and Manolis Mavrikis

ROUTLEDGE

Digital Technology and Artificial Intelligence in Mathematics Education Assessment



erme

european society for research in mathematics education

DIGITAL TECHNOLOGY AND ARTIFICIAL INTELLIGENCE IN MATHEMATICS EDUCATION ASSESSMENT

This timely and thought-provoking book explores how artificial intelligence (AI) and digital technologies more generally are reshaping educational assessment in the field of mathematics. It brings together a rich collection of international perspectives that offer a diverse and critical examination of the opportunities and challenges that digital technologies, including artificial intelligence, play in assessment.

The chapters span a broad range of topics, including the transition to a digital assessment format, designing multiple-choice items, and the use of AI for automated diagnosis, example generation, and feedback provision. The book includes a framing chapter that critically examines current trends in the field and offers key recommendations for rethinking assessment. It calls for more empirical research and a fundamental rethink of assessment practices in light of digital technologies and AI. The volume as a whole also invites readers to engage in a broader conversation about the future of educational assessment and the role of digital technologies in transforming assessment in mathematics education.

This essential resource for educators, researchers, curriculum developers, policymakers, and assessment stakeholders offers a comprehensive look at the evolving international landscape of assessment in mathematics education in the digital era. It also highlights directions for future research and encourages readers to shaping the future of digital technology and artificial intelligence in formative and summative assessment in mathematics education.

Eirini Geraniou is Professor of Mathematics Education at University College London, UK. She has extensive expertise on task design, course design, and design-based research and the integration of digital technologies in mathematics education. Eirini has been an active member of the European Society for Research in Mathematics Education (ERME) since 2009 and became an elected member of the ERME board in 2021.

Cosette Crisan is Professor at University College London, UK. She is a leading expert at the intersection of mathematics education and digital technology. Cosette has taught mathematics at university and school levels, as well as educated future mathematics teachers. She is a member of the prestigious London Mathematical Society Education Committee, which is involved in a wide range of activities, including supporting mathematics education in schools, colleges, and universities.

Manolis Mavrikis is Professor of Artificial Intelligence and Analytics in Education at the UCL Knowledge Lab – an interdisciplinary research centre at UCL Institute of Education. A former Turing fellow at UK's Alan Turing Institute (ATI) and director of the master's degree in education and technology, where he continues to teach and supervise students in artificial intelligence in education, Manolis has led large national and international projects in educational technology.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

New Perspectives on Research in Mathematics Education – ERME series

The European Society for Research in Mathematics Education (ERME) represents a vibrant and growing community of approximately 2500 researchers from across Europe and beyond. United by a shared commitment to understanding and improving the teaching and learning of mathematics, ERME fosters collaboration across borders, disciplines, and educational contexts.

The *ERME Book Series*, published by Routledge, is a key vehicle for documenting and disseminating the substantial and diverse body of research emerging from this community. Volumes in the series include scholarly monographs and edited collections that grow out of sustained collaborations among European researchers in mathematics education. While firmly rooted in the European tradition, the series is international in scope – written by and for researchers not only within Europe, but also around the world.

A hallmark of the series is its commitment to rigorous academic standards. Each volume undergoes a thorough and constructive peer review process, overseen by an international advisory board that ensures both scholarly excellence and global relevance. This process helps maintain the quality and integrity of the series, while encouraging innovative and critical scholarship.

The contributions featured in the *ERME Book Series* aim to deepen our understanding of mathematics education across a wide range of settings:

from schools and universities to informal learning environments. A particular emphasis is placed on examining the institutional, societal, and cultural contexts that shape how mathematics is taught, learned, and researched. The series embraces a plurality of perspectives, theoretical frameworks, and methodological approaches, reflecting the richness and diversity of the field.

By fostering wide authorship and encouraging dialogue across contexts, the *ERME Book Series* contributes to a shared, yet multifaceted vision of mathematics education research – one that continues to evolve in response to the needs and challenges of our time.

Volumes of the series:

(1) Developing Research in Mathematics Education
Twenty Years of Communication, Cooperation and Collaboration in Europe
Edited by Tommy Dreyfus, Michéle Artigue, Despina Potati, Susanne Prediger, and Ken Ruthven

(2) Working with the Anthropological Theory of the Didactic in Mathematics Education
A comprehensive casebook
Edited by Marianna Bosch, Yves Chevallard, F. Javier García, and John Monaghan

(3) Professional Development and Knowledge of Mathematics Teachers
Edited by Stefan Zehetmeier, Despina Potari, and Miguel Ribeiro

(4) Classroom Research on Mathematics and Language
Seeing Learners and Teachers Differently
Edited by Núria Planas, Candia Morgan, and Marcus Schütte

(5) Research and Development in University Mathematics Education
Overview Produced by the International Network for Research in Didactics of University Mathematics
Edited by Viviane Durand-Guerrier, Reinhard Hochmuth, Elena Nardi, and Carl Winsløw

(6) Mathematics Education in the Digital Age
Learning, Practice and Theory
Edited by Alison Clark-Wilson, Ana Donevska-Todorova, Eleonora Faggiano, Jana Trgalova, and Hans-Georg Weigand

For more information about this series, please visit: www.routledge.com/European-Research-in-Mathematics-Education/book-series/ERME.

DIGITAL TECHNOLOGY AND ARTIFICIAL INTELLIGENCE IN MATHEMATICS EDUCATION ASSESSMENT

Edited by Eirini Geraniou, Cosette Crisan, and Manolis Mavrikis

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Cover image: ViewStock via Getty Images

First published 2026
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-032-87653-5 (hbk)
ISBN: 978-1-032-87652-8 (pbk)
ISBN: 978-1-003-53376-4 (ebk)

DOI: 10.4324/9781003533764

Typeset in Optima
by Apex CoVantage, LLC

CONTENTS

ABOUT THE EDITORS

Professor Eirini Geraniou (UCL) has extensive expertise on task design, course design, and design-based research and the integration of digital technologies in mathematics education. Eirini has been an active member of the European Society for Research in Mathematics Education (ERME) since 2009 and became an elected member of the ERME board in 2021. She has received a UCL Faculty Education Award (April 2021) for outstanding contributions to the learning experience and success of students. She has worked in many national and international research projects that focused on the design and evaluation of innovative educational digital resources and media and the assessment of students' learning outcomes during and after their interactions with digital resources. She is also one of the editors of the *Research in Mathematics Education* journal, a leading peer-reviewed journal for mathematics education research.

Eirini Geraniou Research and Publications Profile –
https://profiles.ucl.ac.uk/48410-eirini-geraniou

Professor Cosette Crisan (UCL) is a leading expert at the intersection of mathematics education and digital technology. With a strong track record of integrating digital technology across all levels of education, Cosette has taught mathematics at university and school levels as well as educated future mathematics teachers. Her research focuses on subject-specific knowledge within the school curriculum and teacher education, including subject-specific mentoring. Notably, she demonstrated initiative and leadership in designing the online module 'Digital Technology for Mathematical Learning' for UCL's internationally renowned masters' in mathematics education programme. This module draws on her extensive research experience and expertise in using digital technologies to enhance conceptual understanding of mathematics, informed by her years of collaborative work with teachers and students.

Cosette Crisan Research and Publications Profile – https://profiles.ucl.ac.uk/48434-cosette-crisan

Professor Manolis Mavrikis (UCL) is based at the UCL Knowledge Lab – an interdisciplinary research centre at UCL. A former Turing fellow at UK's Alan Turing Institute (ATI) and director of the master's degree in Education and Technology, he continues to teach and supervise students in artificial intelligence in education. Beyond his publications, he has contributed to the field through various funded projects and partnerships with schools, governments, and third-sector organisations. Notably, Manolis led the design and evaluation of the adaptivity and feedback of educational platforms under large EU-funded projects, such as iTalk2Learn and iRead, and UK-funded projects, such as Designing and Defining Children's Agency in the Age of AI (CHAILD). He is also one of the editors of the *British Journal of Educational Technology*, a leading peer-reviewed journal in the field.

Manolis Mavrikis Research and Publications Profile – https://profiles.ucl.ac.uk/48731-manolis-mavrikis

CONTRIBUTORS

Hassan Ayoob University of Haifa, Israel

Sami Baral Worcester Polytechnic Institute, USA

Bärbel Barzel University of Duisburg-Essen, Germany

Rogier Bos Utrecht University, Netherlands

Anthony F. Botelho University of Florida, USA

Mats Brunström Karlstad University, Sweden

Li Cheng University of North Texas, USA

Francesco Contel Sapienza University of Rome, Italy

Cosette Crisan University College London, UK

Annalisa Cusi Sapienza University of Rome, Italy

Paul Drijvers Utrecht University, Netherlands

Maria Fahlgren Karlstad University, Sweden

Eirini Geraniou University College London, UK

Ramona Hagenkötter University of Duisburg-Essen, Germany

Bastiaan Heeren* Open University, Netherlands

Neil T. Heffernan Worcester Polytechnic Institute, USA

* *We are very sorry that Chapter 11's co-author, Bastiaan Heeren, passed away before the publication of this book.*

Paola Iannone University of Edinburgh, UK

Uffe Thomas Jankvist Aarhus University, Denmark

Johan Jeuring Utrecht University, Netherlands

George Kinnear University of Edinburgh, UK

Katrin Klingbeil University of Duisburg-Essen, Germany

Alice Lemmo University of L'Aquila, Italy

Maria Margeti University College London, UK

Manolis Mavrikis University College London, UK

Morten Misfeldt University of Copenhagen, Denmark

Filip Moons Utrecht University, Netherlands

Shai Olsher University of Haifa, Israel

Michael O. Oyengo Maseno University, Kenya

Guido Pinkernell University of Education Heidelberg, Germany

Mary Richardson University College London, UK

Fabian Rösken University of Duisburg-Essen, Germany

Christopher Sangwin University of Edinburgh, UK

Samuel Sollerman Stockholm University, Sweden

Anica Stemmer University of Duisburg-Essen, Germany

Marianne Thomsen University College Copenhagen, Denmark

Paul Tyrichter University of Duisburg-Essen, Germany

Gerben van der Hoek Utrecht University, Netherlands

Hendrik Van Steenbrugge Stockholm University, Sweden

Thomas Vikberg University of Helsinki, Finland

Hans-Georg Weigand University of Würzburg, Germany

Mattias Winnberg Stockholm University, Sweden

FOREWORD

In recent decades, our world seems to be transformed on a daily, perhaps hourly, basis by new technologies claiming a vast range of ways to impact our lives, from improving efficiency to lightening our workloads, reducing complex decision-making tasks, diagnosing illnesses, writing volumes of text, and more. Much more. Indeed, as the authors of this book demonstrate, our world as we know it is being transformed by technologies, many of which incorporate forms of artificial intelligence (AI), and we are increasingly reliant on such technologies for mass communication, aspects of employment, entertainment, and the management of our daily lives.

This book presents a highly focused view of assessing one subject, mathematics, and the chapter authors discuss a surprisingly wide range of technologies, including interactive environments, computer-aided assessment, and automated approaches to formative methods, such as feedback. Some important questions need to be asked continuously about assessment, not only how we might assess knowledge and understanding of any subject in a digital era, but also what we should assess and, most importantly, why. Instead of debating whether technology, including AI, should be incorporated more into formative and summative assessments, we should perhaps be asking why this would not be the case.

However, when we step into the world of education, the pace of change is suddenly slower and it appears that much of our practice is almost untouched by new technologies, and this is evident in mathematics, where digital tools may influence teaching and learning on occasions, but substantive change is rare. Peer even closer, into the realm of educational assessment, and what we do to students reveals assessment frameworks that are similar to those seen some 100 years ago. Our favoured approaches to teaching and assessing still employ what can be termed traditional, summative assessments and the privileging of just a few very high-stakes assessments (again, mainly tests), and most still delivered on paper. It seems that clinging to tradition or relying on general inertia coupled with an aversion to any changes that might appear risky is a successful means of maintaining the (technologically sparse) status quo in assessment.

It is clear that we need to move from an idea where digital assessment is simply the paper design and items sit behind glass; the past is helpful insofar as providing an understanding of the immense body of research that now exists around assessment, but the future is going to require significant vision and courage to experiment. In a time when disadvantage and its implications for educational outcomes continue to make headline news around the world, perhaps what matters most, in terms of introduction and use of digital technologies to support and enhance assessment, is equity. Globally, we know that access to digital resources for education is very unequal, and even in wealthy countries such as England, there are digital divides that prevent all learners from accessing the most up-to-date technologies for teaching and learning.

Using mathematics as a frame of reference allows this international collection of authors the opportunity to interrogate some of the difficult questions around how technology challenges us to investigate what is meant by learning mathematics with technology and how assessment can show that mathematical learning has happened. The assessment landscape is constantly changing, because what we learn from research demonstrates the need for human responsiveness to the shifting nature of education itself. We cannot, nor should we, rely on 'one way' to undertake assessment of learning, because it is the most valuable aspect of education and of being educated. It is also important to be realistic about how fast or slowly change can, or should, be enacted; in recent years, the advent of large language models (LLMs) has unleashed something of a moral panic across assessment research communities and within the assessment industry. But in the fog of fear about potential massification of cheating in assessment, we must be careful not to lose sight of the affordances that new technologies can bring to educational settings. As the introductory chapter by the editors suggests, the same advances in AI can facilitate approaches that include responsive and adaptive assessments that are integral to the professional role of teachers. As such, the authors present ways in which such tools can be responsibly used to support learning, teaching, and assessment. What this book presents is a balanced view with ideas and examples that have a global applicability regardless of subject expertise, because it is assessment that is emphasised in each chapter.

There can be no perfect answer to choosing a 'solution' in terms of assessment practice, because assessment (whether its genesis, and enactment, is digital or on paper) is an imperfect practice. What really matters is developing an evidence-based approach to good practice and ensuring that we remain curious, because to paraphrase Dylan Wiliam's words from 2018, when asked 'what works', we must stop seeking a silver bullet because, in assessment, just about everything works somewhere, but nothing works everywhere.

Mary Richardson
Professor of Educational Assessment

1

SETTING THE SCENE ON THE USE OF TECHNOLOGIES AND ARTIFICIAL INTELLIGENCE IN FEEDBACK AND ASSESSMENT IN MATHEMATICS EDUCATION

Eirini Geraniou, Cosette Crisan, and Manolis Mavrikis

1 Introduction

Assessment is a crucial element in any learning process, and living in the digital era, we cannot ignore the influence and impact that digital technologies (DT) have in mathematics education. Using DT in formative and summative assessment in mathematics provides numerous new possibilities in terms of ways to assess mathematical learning but undoubtedly also leads to new challenges (e.g. Jankvist et al., 2021; Drijvers & Sinclair, 2023). For example, it has become easy to assess students' procedural skills in a digital environment, while on the other hand, assessing one's 'working-out' can be extremely problematic, as digital environments often fail to capture the step-by-step reasoning behind an answer.

In the last 15 years, there have been arguments about how assessment practices and policies in most educational systems have been 'slow' to transform or adapt in light of advances in DT (e.g. Shute & Becker, 2010; Timmis et al., 2016). There is a relatively small body of work which has begun to explore how DT, including the more recent artificial intelligence (AI) tools, offer opportunities for innovation, as well as risks and challenges for rethinking feedback and assessment purposes in mathematics education (Stacey & Wiliam, 2013; Aldon & Trgalová, 2019; Clark-Wilson et al., 2021). In fact, a few years ago, Drijvers (2018) foretold that 'digital assessment of mathematics is a phenomenon that will play an increasingly important role in mathematics education' (p. 63). Indeed, how to design assessments with and through digital technology (Drijvers, 2018; Stacey & Wiliam, 2013)[1] that effectively evaluate students' mathematical knowledge and understanding, rather than their digital competencies, is an under-researched area. Current attempts in

DOI: 10.4324/9781003533764-1

integrating DT into assessment of mathematics rarely go beyond digitisation of the assessment process, where the approaches to assessing mathematics are the same as they have always been, apart from assessment 'happening' via a digital tool. More recently, researchers have turned their attention towards how digital technology could be integrated into assessment to support students doing mathematics, and to express themselves mathematically, and to show their working-out, cf. an overview of how technology has been transforming teaching, learning, and assessment in mathematics education in the digital age in Weigand et al. (2024).

To set the scene for our discussion of the book's contributions and its individual chapters, it is useful first to clarify the key terms used when discussing digital technology. We and the authors of the various chapters take a broad spectrum of tools ranging from dynamic geometry environments to computer algebra systems and from automated assessment platforms to AI-based applications. The chapters in this book discuss various uses of these technologies – from interactive environments that allow mathematical manipulation and construction to computer-aided assessment systems that automate feedback and grading. When considering specifically AI, we take an inclusive perspective that spans both recent developments in large language models (LLMs) and also established approaches that are rooted in more traditional AI techniques, such as symbolic approaches that are used in computer algebra systems (CAS).

This book explores several key aspects, such as digital resources for different types of assessment, AI in/for assessment in mathematics, benefits and challenges of using digital technology for assessment, international perspectives on using digital technology for assessment, and innovative approaches to technology-enriched assessments in mathematics. Before analysing the specific contributions of the chapters and the book as a whole in Section 3, we present an overview of the 14 chapters in Section 2 to help the reader understand the broad landscape of the issues, approaches, and contexts addressed. Rather than attempting a comprehensive literature review, we have highlighted some key issues brought forward in each chapter and point the reader to those for more detailed literature reviews. We conclude by advocating the need for mathematical assessment reform when DT and AI are involved and a research-informed future agenda for mathematical assessment in the digital era.

2 Overview of the chapters and key issues

In **Chapter 2, Paola Iannone, Alice Lemmo, and George Kinnear** put forward and discuss several issues with regards to the transition from traditional paper-based assessments to e-assessment. The authors look into literature regarding the characteristics of e-tasks, but also transfer of assessment tasks to e-mode, to establish a foundation for analysing the shift to e-assessment and how this impacts students' learning and assessment validity. Employing the instrumental approach to analyse these effects is a valuable contribution, providing a theoretical perspective that enriches the empirical findings observed in the case studies they present. These are two examples of task transfer from paper-based to e-assessments, one being at middle school level (grade 6) in Italy and one at university level (first year maths undergraduate students) in the UK. The authors argue that e-assessment allows for the assessment of skills that are often different than those assessed in the paper-based formats, since students use different schemes (or in other words, different ways of working) in either mode of assessment. The authors also address broader implications for educational policy and assessment, particularly the need for further research and appropriate evidence in ensuring reliability and validity when interpreting scores from e-assessed learning and, particularly, in large-scale national assessments. They also discuss potential challenges to some of the issues identified in the chapter, for example, how to strategically use different assessment modes and design tasks to target specific learning objectives.

In **Chapter 3**, **Mattias Winnberg, Samuel Sollerman, and Hendrik Van Steenbrugge** provide an insightful exploration of the assessment of students' mathematical knowledge through digital items in large-scale assessments and how the learning opportunities in the introduction to these items connect to the technical and conceptual understandings required to solve the assessment items. The study is based on the PISA 2022 international large-scale assessment, which evaluates students' mathematics knowledge globally. The chapter contributes to understanding how digital tools, such as spreadsheet-like tools, are used to assess the mathematical problem-solving skill of students in secondary education, and the validity issues surrounding their use. Recognising that the focus of their study is confined to the items themselves, Winnberg et al. propose that future research is needed to explore further the validity of using digital tools in large-scale assessments, particularly regarding how students' technical and conceptual understanding is evaluated through spreadsheet items, and investigate how these tools affect the development of students' cognitive and problem-solving abilities.

Chapter 4, by **Thomas Vikberg**, focuses on the transition from handwritten to digital large-scale mathematics assessments, exploring the actors, political context, and historical timeline behind this shift. The author provides a detailed case study of Finland's approach to digitalisation of high-stakes mathematics assessment, with a particular focus on large-scale assessments in the context of mathematics education at general upper secondary school level (Matriculation Examination). The chapter contributes valuable insights into the political and historical aspects of the digital transformation in mathematics assessments. Given the large volume of submitted digital test answers over the past five years since mathematics assessment transitioned to a digital format in 2019, the author highlights the potential of conducting further research into investigating the digitisation of high-stakes examinations, which can also serve as a model for other educational systems, particularly in the context of open-ended assessment items.

In **Chapter 5**, **Hassan Ayoob and Shai Olsher** compare the use of digital formative assessment (DFA) tasks and traditional assessment and discuss the impact of using DFA tasks on teachers' ability to assess individual students and the entire class in elementary school geometry. They carried out a study with nine elementary school teachers teaching grade 5 students in Israel. The DFA tasks were designed to assess different aspects of students' answers in geometry, revealing their strengths and abilities, as well as their difficulties and their needs. In addition, teachers' dashboards provided teachers with results (on both the student and class levels) to facilitate formative assessment practices. Their findings reveal that teachers recognise the value of using DFA. They are automatically assessed, and the results are immediately available and accessible for single students, as well as for the entire class, providing teachers reports in offering immediate, specific, and actionable feedback. The findings presented emphasise the effectiveness of DFA tasks in elementary geometry, especially for dynamic and visual concepts like triangle altitudes and polygon areas. However, its focus on geometry restricts applicability to other areas, such as algebra, where interactive visuals may not be as beneficial or widely used. While the teachers viewed the tasks positively, the lack of qualitative data limited a deeper understanding of their feedback. Additionally, although DFA tasks offer immediate feedback and structured progression, they might not be as adaptable as traditional formative assessments to meet various classroom needs. The authors highlight that future research should investigate the use of DFA in different subjects, and how teacher training affects its implementation.

In **Chapter 6, Guido Pinkernell and Hans-Georg Weigand** investigate the formative assessment of *conceptual knowledge*. They recommend the need for assessment tasks that allow learners to show their conceptual knowledge, and look into this aspect considering the specific contributions of technologies in the digitisation of assessment (e.g. dynamisation, interaction, and multiple representation of mathematical concepts). Their contribution lies in discussions about task and feedback design and in presentations of examples of how conceptual knowledge could be assessed and supported in a digital environment in a dynamic, interactive, and multi-presentational way. They suggest focusing on basic mental models (BMMs), which can be considered as the central core of understanding of mathematical concepts. But they acknowledge the lengthy and demanding theoretical and empirical process required in developing BMMs, followed by the necessary development of adequate tasks. Also, they argue about the importance of feedback (including adaptive) and its design for formative digital assessments. The authors call for the need to develop and empirically validate the development and use of a digital environment with content tasks that formatively assess conceptual knowledge, while being easily accessible by learners and practical for teachers.

In **Chapter 7**, **Sami Baral, Li Cheng, Anthony F. Botelho, and Neil T. Heffernan** share the technical innovations in automated assessment for open-ended mathematics questions. Noting that open-ended questions are important in assessing deeper understanding and critical thinking, the authors suggest that automated approaches could support teachers and students particularly in resource-constrained environments. The chapter presents the authors' approach, which combines sentence-level semantic representations with a specialised mathematical terms model that improves assessment accuracy for responses that contain both text and mathematical expressions. The authors identify areas for development, including the need for detailed feedback, but also potential issues with the approach, such as fairness and bias, student privacy, and personal information.

In **Chapter 8**, **Annalisa Cusi and Francesco Contel** analyse upper secondary students' interactions with GPT-4o for self-assessment in mathematics. Employing instrumental genesis theory as a lens to understand how students develop their use of AI tools over time, the authors focus on the potential role technology can play as a metacognitive tutor. Through a study of 19 students working with GPT-4o on mathematical problem-solving tasks, the authors identify several key findings about how students perceive and develop their use of AI as a tutoring tool. In particular, they propose reconceptualising instrumental genesis through (1) the impact of humanising AI and users' metacognitive skills, (2) the role of uncertainty and randomness in AI interactions, (3) the indeterminate nature of user–AI co-actions, and (4) the unstable nature of utilisation schemes with AI tools. The authors advocate how this instability is fundamentally different from how students develop stable patterns of use with other educational technologies and use the metaphor of 'noise reduction' from physics to describe how users must develop metacognitive strategies to navigate this uncertainty of interactions with LLMs while working toward productive learning outcomes.

In **Chapter 9**, **Bärbel Barzel, Ramona Hagenkötter, Katrin Klingbeil, Fabian Rösken, Anica Stemmer, and Paul Tyrichter** highlight the value of using multiple-choice (MC) items in formative and summative assessments, as MC items can be quickly administered, and students' answers automatically analysed, allowing a quick insight into students' current understanding. They draw on scientific findings from both (educational) psychology and mathematics education to offer guidelines on how to design and validate MC items, using an example from the digital formative assessment tool SMART. Additionally, various scenarios are presented to show how carefully designed MC items, particularly when integrated with digital technology, can be used both to assess and to improve students' understanding, and to give teachers valuable insights into potential misconceptions and key aspects of a mathematical concept, thus strengthening their pedagogical content knowledge. The authors call for further research on the particular role MC items in digital formative assessments can play in mathematics education. They suggest that involving students in creating and evaluating formative (and even summative) MC assessments is promising for nurturing mathematics teaching and learning.

Chapter 10, by **Maria Fahlgren and Mats Brunström**, explores how the integration of digital tools, specifically computer-assisted assessment (CAA) systems, could support student learning in mathematics. More precisely, it highlights the potential of these tools by focusing on a specific type of task: example-generation tasks (suitable for the combined use of a dynamic mathematics software [DMS] and a CAA system), where students are prompted to create examples that meet certain conditions, as a means to actively engage students in developing a deeper understanding of mathematics. The chapter examines the complexity of designing example-generation tasks, and the authors call for further research on how various design choices within CAA systems may enable students to enrich their example spaces and apply key ideas that effectively foster a rich and varied example space while addressing key mathematical ideas. Although the research took place at a Swedish university involving three groups of first year engineering students enrolled in an introductory calculus course, the findings of this study are broadly applicable beyond Sweden.

In **Chapter 11, Gerben van der Hoek, Bastiaan Heeren, Rogier Bos, Paul Drijvers, and Johan Jeuring** examine students' use of a balanced feedback strategy that involves balancing between feedback that allows for exploration and feedback that mitigates learning barriers and also look into students' opinions about feedback. They carried out a study in the Netherlands with 25 15- to 17-year-old senior general secondary education students who practiced linear and exponential extrapolation in an online environment that offered feedback following the previously described balanced feedback strategy. Their findings reveal that such a balanced feedback strategy can be effective in an online learning environment as it promotes fruitful student–environment interactions enabling exploration and that the students appreciate balanced feedback. They call for further research on how such a balanced, informative tutoring feedback strategy would perform in comparison to a feedback strategy that, for example, provides worked-out examples and knowledge of results.

In **Chapter 12, Christopher Sangwin and Michael O. Oyengo** examine how to move beyond simple right/wrong evaluations towards assessing higher-order mathematical thinking. They consider how student-constructed mathematical expressions can be assessed in automated assessment systems in two case studies in Scotland and Kenya. The chapter provides concrete strategies for implementing assessment tasks that leverage a computer algebra system (CAS) that can verify the mathematical properties of students' answers. The answers must satisfy certain conditions that an author specifies, allowing this way a rigorous verification of mathematical properties. The authors also demonstrate how shared digitised courses can benefit a developing country such as Kenya. The chapter highlights how different technical approaches serve different assessment purposes – while modern machine learning and generative AI approaches may excel at natural language processing, a CAS provides precise verification of mathematical properties and relationships, and therefore, they believe that CAS-based assessment will remain valuable alongside AI tools, with each technology finding its appropriate niche based on effectiveness rather than novelty. The authors recommend that institutions provide technical specialists to support teachers in authoring and suggest that AI approaches could enhance feedback quality by analysing large datasets of student interactions to improve question design.

Chapter 13, by **Maria Margeti and Manolis Mavrikis**, presents students' engagement with interactive learning environments (ILEs) and how these tools support their learning through formative assessment. Using Entwistle's theory of deep and surface learning approaches (Entwistle, 2001) as a conceptual framework, the study explores the interplay between these approaches among first year undergraduate students at a university in the UK. By analysing interaction data within ILEs, the research demonstrates how students' learning approaches can be identified and evaluated. The authors suggest that future similar research could refine the identification of learning approaches within ILEs, enabling tailored interventions to individual learning patterns. Such insights could inform the enhancement of ILE design features, further supporting meaningful formative assessment and improving educational outcomes. Furthermore, the authors propose that using machine learning approaches to analyse patterns in student interaction data could further support educators but requires recognising that effectively interpreting and acting upon the learning analytics data that these ILEs provide require further professional development.

Chapter 14, by **Filip Moons**, introduces an approach to bridging human and automated formative assessment through 'atomic feedback' – re-usable, discrete units of focused commentary. This is a 'human in the loop' approach where technology supports rather than replaces teacher judgement. The chapter shows that when teachers use atomic feedback principles, they provide more detailed feedback, though this doesn't automatically translate to higher-quality feedback. Looking to the future, the author propose integrating with other AI approaches that can analyse student interactions and preferences and make more individualised feedback suggestions while maintaining the crucial role of teacher judgement. The work has particular implications for assessment transparency and student understanding, with evidence suggesting that struggling students especially benefit from the clear, structured feedback this approach provides.

In **Chapter 15, Marianne Thomsen, Morten Misfeldt, and Uffe Thomas Jankvist** make a theoretical contribution by examining AI literacy through the lens of mathematical competencies and different forms of mediation. Drawing on Misfeldt and Jankvist's (2018) framework of epistemic, pragmatic, and justificational mediations, the chapter explores how 'techno-authoritative conviction schemes' (as the authors term them) might lead students to uncritically accept AI-generated mathematical explanations. The authors provide an example that uses historical sources alongside AI tools to encourage critical engagement and recommend designing open assessment situations that require students to actively question and evaluate AI-generated responses. The chapter highlights key challenges for teachers given the variable nature of AI responses and emphasises the importance of moving beyond purely justificational uses of AI.

3 What this book contributes

The overview of each chapter presented in the previous section reflects the diverse theoretical and practical perspectives and efforts from the international community of mathematics educators and researchers contributing to this book as they explore the future of DT and AI. In terms of theoretical underpinning, chapters range widely from employing microdidactic frameworks rooted in cognitive approaches (such as instrumental genesis) to more 'macro' perspectives, such as competency-based frameworks. This diversity reflects the nature of the chapters when considering how the book contributes beyond simply reporting innovation but towards informing educational practice and policy. In doing so, the book raises important questions about the assumptions, goals, and future directions of research in this area. We call on readers to consider the following fundamental questions that emerge across the chapters regarding what we assess, how we assess it, and why:

- How can we assess knowledge and understanding in the digital era, and should we?
- Are we able to identify the evidence for learning? What does *learning* mean in the digital era? Does it mean knowing what a mathematical concept is? Knowing how to use DT and AI to solve a mathematical problem? Being able to explain and justify a solution and an answer?
- How can we 'see' and assess the trail of learning a student leaves when interacting with DT and AI?
- What are the design principles for assessment tasks presented via the use, or involving the use, of DT?
- Do teachers have the necessary skills (digital and data literacy) to assess learning, knowledge, and understanding?

Specifically with regard to AI:

- How can we ensure that AI-supported feedback and assessment follows a balanced approach to both conceptual knowledge and procedural fluency evaluation?
- What is the appropriate balance between automation and human judgement in AI-supported assessment?
- What new forms of feedback and assessment might become possible through AI, and how do we ensure these align with educational goals?

The aspects related to the preceding questions have been discussed in various chapters in this book, and we, as authors of this first chapter and editors of this book, have identified five categories which we believe capture key issues necessary to answer these questions. These are (1) transitioning from traditional to digital assessment, (2) assessing deeper understanding and conceptual knowledge in digital contexts, (3) enhancing assessment through DT, (4) task design for digital and AI-based assessment, and (5) human–AI partnerships. In what follows, we discuss each chapter's contributions against each category.

3.1 Transitioning from traditional to digital assessment

The transition of mathematics assessments to digitised formats comes with its own set of challenges. Writing mathematics in a digital environment is far from straightforward, unless aided by tools like smart pens. Even then, some claim that the act of *doing* mathematics in digital environments is fundamentally altered, as it is less effortful in comparison with pen and paper (Aspiranti et al., 2020). Something essential seems to be missing. On paper, mathematical expressions, equations, graphs, and drawings naturally allow for a variety of informal annotations – scribbles, strikethroughs, erasures, arrows linking ideas, circling, and underlining – all of which play a crucial role in the process of thinking and doing mathematics. These markings on the proverbial mathematician's scribbles on a napkin at a party make the reasoning visible, capturing some of the flow between the thinking and the doing of mathematics. Such spontaneous, flexible interactions are not always easily replicated in a digital environment, potentially obscuring evidence of mathematical thinking and problem-solving. This is a problem that had to be overcome following the digitalisation of Finland's Matriculation Examination, as described in **Chapter 4**, for example. The author explains how the initial approach to address concerns with writing mathematical notation in such tests was to identify an existing formula editor suitable for use by students during the exams. However, the lack of a suitable editor and the constraints of the development timeline led to the decision to design a new editor, specialised, exam-specific tool which also allowed inclusion of screen captures from multiple mathematics software tools in the answers, such as tools for creating images and diagrams, for manipulating spreadsheet data, and function GraphWare.

While some aspects of the mathematical experience may be diminished in digital spaces, the opportunities for new, powerful ways of exploring mathematical ideas present a compelling counterbalance. Consider the example of sketching graphs on paper, which traditionally involves several steps: Draw a table of values; choose values of *x*, then calculate values of *y*; plot (x,y) points on a set of axis; and finally join up the plotted points. Without any doubt, sketching graphs is a skill needed by the student, and such activities are needed if the aim is to provide an increased understanding that a graph consists of points that all meet a given condition. However, the 'doing of maths' is procedural, without developing an understanding of how the equation and the numbers that appear in it determine the shape of the graph (or the other way around). This approach, though valuable, does not necessarily support an understanding of how the equation itself shapes the graph, or how the graph informs our understanding of the equation. In contrast, this is possible when a GraphWare is used: The sketching of the graphs is all done in one click. The digital tool offers different engagement with maths, leading to a different maths learning through discovering properties of functions by quickly constructing a large number of features that can be compared. Students receive instant feedback, allowing them to refine their responses, explore different approaches, and deepen their engagement with mathematical concepts. In this digital environment, the keyboard and the mouse replaced the pen, the screen replaced the paper in the notebook, the on-screen equation editors and graphical packages replaced the pen and geometrical instruments, but most importantly, digital tools shift the nature of the mathematical activity itself, moving away from the 'typical' textbook exercises to more dynamic, exploratory learning.

This transformation also extends to assessment, where the implementation of digital tools offers new ways to measure and support student progress. **Chapter 2** focuses on case studies of transfer from traditional, paper-based tasks to e-tasks, including one at the middle school level and one at the university level. The authors address key challenges and implications for learning and assessment validity when adapting assessment tasks to digital formats. Drawing on literature about the characteristics of e-tasks and their transfer to digital formats, they establish a foundation for analysing this shift. Continuing this focus, **Chapter 3** delves into the transition from traditional, paper-based assessments to digital interactive stimuli (DIS) in large-scale evaluations, with a focus on assessing students' mathematical problem-solving abilities. DISs are particular types of digital items, with characteristics and functions that are improbable to include in paper-based assessments, containing interactive stimuli, such as spreadsheet, graphing, dynamic geometry, or computer algebra systems (CAS) tools.

Equally important to the shift to digital assessment, **Chapter 4** provides insights into the systemic and policy-driven transitions necessary for digitisation, drawing lessons from Finland's approach to transforming high-stakes assessments.

3.2 Assessing deeper understanding and conceptual knowledge in digital contexts

In the mathematics education literature, there is a distinction between two types of 'understanding', that is, Skemp's (1976) work on relational/instrumental understanding, and two types of 'knowledge', that is, Hiebert and Lefevre's (1986) work on conceptual/procedural knowledge. When reflecting upon the conceptualisation process any student goes through, we need to consider how they acquire mathematical knowledge and understanding and what it means that they acquire these.

While DT offer new ways to assess what students know and understand, how students approach their learning remains equally important. Entwistle's work (e.g. Entwistle, 2001) on approaches to learning lies in how learners engage with material at different levels, shaping their understanding and knowledge construction. According to Entwistle, deep approaches to learning involve a genuine intention to understand the material. On the other hand, surface approaches to learning are characterised by rote memorisation and rote learning, following procedures with not much understanding, which can often lead to cognitive gaps or fragmented understanding. By using Entwistle's theory, educators can better design interventions that encourage desirable deep learning and meaningful engagement with concepts. This is evident in **Chapter 13**, where the authors illustrate how surface-level learning intentions can be identified through log data in interactive learning environments (ILEs), opening up the possibility of using ILEs to provide targeted tutor guidance to encourage deeper approaches to learning.

Similarly, using PISA 2022 as a case study, **Chapter 3** explores the integration of spreadsheet-like tools, emphasising the implications for evaluating both 'technical and conceptual understanding' by building on the reasoning of Drijvers and Gravemeijer (2005), who stress the interconnected and co-evolving relationship between techniques and conceptual understanding when digital tools are incorporated into mathematical problem-solving activities.

We appreciate the complexity of assessing students' relational understanding, as defined by Skemp (1976), as well as their conceptual knowledge (as defined by Hiebert and Lefevre, 1986) in a digital environment, or even with the help of DT. However, if we are to argue about a successful integration of DT in mathematics education and for assessing mathematics learning, we must look into how students' conceptual knowledge and relational understanding of mathematical concepts, as well as their problem-solving skills, can be assessed using DT. We therefore need to research further how DT influence the development of students' cognitive and problem-solving abilities, as also recommended by the authors of **Chapter 3**.

3.3 Enhancing assessment through digital tools

A number of chapters focussed on the potential of digital assessments that go beyond procedural knowledge to foster higher-order thinking and conceptual knowledge in mathematics. This could be achieved in various ways: example-generation tasks with computer-assisted tools like GeoGebra and Mobius that can encourage deeper mathematical exploration (**Chapter 10**), computer algebra systems (CAS) that enable focus on mathematical concepts rather than calculation (**Chapter 12**), spreadsheet items in PISA assessments (**Chapter 3**), or other interactive environments that allow students to construct and manipulate mathematical objects (**Chapter 5**). More specifically, **Chapter 10** explores how encouraging students to create their own examples within a computer-assisted system supports their conceptual depth but also can illuminate certain misconceptions. **Chapter 12** showcases how CAS can shift students' attention from carrying out algorithms to understanding underlying mathematical structures, while **Chapter 3** unpacks large-scale assessments that employ spreadsheet-like tools to test higher-order problem-solving skills.

The provision of feedback is also an important aspect of enhancing assessment through digital tools. Computer algebra systems, for example, can provide immediate feedback (**Chapter 12**), and similarly, **Chapter 8** demonstrates the potential of AI, such as ChatGPT, as a metacognitive mathematics digital 'tutor'. Efforts to automate the assessment of open-ended questions show promise in providing more sophisticated feedback mechanisms (**Chapter 7**). The challenge of providing meaningful feedback at scale is addressed through various approaches, from semi-automated systems that combine human expertise (**Chapter 14**) to adaptive feedback systems designed to support conceptual understanding (**Chapter 6**). These attempts to enhance assessment through digital tools have important implications about the nature of understanding and how it can be assessed. Multiple-choice items, when carefully designed, can diagnose conceptual understanding rather than just test recall (**Chapter 9**). The authors of **Chapter 9** illustrate how choosing the right distractors and integrating them within a digital tool can help identify and address persistent misconceptions more effectively. Similarly, example-generating tasks can reveal students' deeper grasp of mathematical concepts by requiring them to construct their own mathematical objects meeting specific criteria (**Chapter 10**). These approaches demonstrate how digital tools can support more sophisticated forms of assessment that go beyond traditional procedural checks.

All in all, the chapters under this category demonstrate that careful attention is needed in both pedagogical principles and the ways technology can be leveraged to genuinely enhance feedback and assessment. Central to achieving this enhancement is the thoughtful design of assessment tasks, which is the focus of the next category.

3.4 Task design focusing on assessment involving the use of digital technologies and AI

Different technological tools enable distinct types of activities, each shaping learners' mathematical understanding in unique ways. Undoubtedly, integrating DT in ways that take advantage of these tools' capability to visualise and link symbolic, graphical, and numerical representations and foster greater exploration, inquiry, and collaboration could indeed revolutionise maths education through adopting transformative approaches to the teaching and learning of mathematics in the digital era (cf., Weigand et al., 2024).

However, a key element of the kind of mathematical thinking and learning enabled by digital technology lies in the nature of the tasks that require its use. For instance, using a dynamic geometry environment where students manipulate the vertices of a triangle to observe 'what changes, what stays the same' with respect to the point of intersection of perpendicular bisectors offers a fundamentally different learning experience compared to using the tool to simply graph given functions. But how should student learning in such technology-driven scenarios be assessed? Drijvers (2018) concluded that 'digital assessment of mathematics is a phenomenon that will play an increasingly important role in mathematics education' (p. 63). Yet the challenge of designing assessments that effectively measure students' mathematical knowledge, rather than their DT skills, remains an under-explored area, as also argued earlier. Current efforts to incorporate DT into mathematics assessments tend to rely on multiple-choice (MC) formats, which primarily reflect a digitisation of traditional methods rather than a fundamental shift in assessment practices.

Emerging research highlights reported in this book draw attention to opportunities and challenges of integrating mathematics-specific DT tools into assessments that enable students to 'truly do mathematics' in a digital context – allowing them to express, demonstrate, and create mathematical ideas (Frenken et al., 2022, p. 61). This theme weaves through **Chapters 2, 3, 4, 9, 10,** and **12**, offering a cohesive exploration of the evolving landscape of digital assessments. **Chapter 2** lays the groundwork by reviewing literature on the characteristics of e-tasks and examines the transfer of assessment tasks to e-mode to establish a foundation for analysing the shift to e-assessment and how this impacts students' learning and validity of the assessment. In **Chapter 4**, Vikberg examines what is assessed when DT are incorporated into large-scale assessments and how this impacts the validity of inferences drawn from test scores. This work raises critical questions about the e-task design and the feasibility of fostering a co-evolving development of technical skills and conceptual understanding within the constraints of a single testing moment. This discussion is further extended (**Chapter 9**) by drawing on scientific findings from both (educational) psychology and mathematics education to offer guidelines on how to design and validate multiple-choice (MC) items for formative and summative assessment purposes. MCs require careful design in order to be used both to assess and to improve students' understanding and to give teachers valuable insights into potential misconceptions and key aspects of a mathematical concept, thus strengthening their pedagogical content knowledge. In a similar vein, **Chapter 12** moves the discussion further by examining how student-constructed mathematical expressions can be assessed in automated systems. Further examples of task design approaches that move beyond prioritising basic mathematical skills and a focus on the correctness of the final answers are put forward in **Chapter 10**. The authors propose to employ a combination of two specific types of digital technology, computer-aided assessment (CAA) and dynamic maths systems (DMS), for automated corrections of students' answers to example-generation tasks. These tasks have been redesigned to enrich students' example spaces using the elaborated and adaptive feedback in CAA systems.

We see these efforts as incremental transitions from pen-and-paper assessments to digitised and digital formats, a progression that aligns with Finland's experience, as insightfully described in **Chapter 4**. Drawing on Ripley (2009), Thomas Vikberg (the author of **Chapter 4**) calls for a more transformative approach to task design. Rather than simply transferring existing tasks to a digital environment or giving in to the tendency in the current body of work to envision increasingly sophisticated test items and response methods, Vikberg highlights the untapped potential of digital tools to assess 21st-century skills more effectively. By embracing this transformative perspective, digital assessments can evolve to support skills such as critical thinking, problem-solving, collaboration, and creativity, which are increasingly essential in modern education and beyond.

3.5 Human–AI complementarity

The recent advances in the field of AI in general, and large language models (LLMs) in particular, have resulted in a lot of rhetoric around AI as replacement tool for human judgement or as a tool for automation or efficiency. Similarly, most of the public discourse tends to be concerned with students using AI to cheat in homework or in assessments. However, a strong theme that emerges from the chapters here is a more nuanced view of AI as a technology that is complementary and amplifies human expertise rather than displacing it. This shift is evident in several contributions where the focus lies on how teachers and students can leverage AI systems to shape more effective and adaptive assessment practices. For instance, **Chapter 14** demonstrates how 'atomic feedback' can be harnessed to support a *human-in-the-loop* strategy. Rather than allowing AI to provide all the feedback, teachers retain control over the detail of the comment, while the technology handles repetitive elements. Similarly, **Chapter 8** discusses how an LLM may serve as a metacognitive tutor, prompting students to articulate their reasoning. In both these chapters, the teacher's professional judgement remains central, something that is often neglected in the field, that is, AI tools contribute generative responses, but the human educator moderates and interprets these insights.

Even in automated assessment of open-ended questions (**Chapter 7**), the focus is not on replacing human assessment but on augmenting it through tools that can handle routine aspects of offering summative feedback (such as correct/incorrect answers, scores, or pre-determined statements of mathematical justifications, etc.), while leaving more complex judgements to teachers (such as complex explanations, decisions on follow-up practice, etc.). Similarly, the exploration of AI literacy in mathematics education (**Chapter 15**) suggests that effective assessment requires both AI capabilities and human critical thinking – teachers and students need to develop the ability to assess, challenge, and guide AI-generated outputs rather than accepting them uncritically (Geraniou et al., 2025). Some of the chapters also show how a reconceptualisation of assessment practices might be needed and that theoretical frameworks such as the instrumental genesis framework (**Chapter 8**) and the mediations framework (**Chapter 15**) show how traditional theoretical frameworks need adaptation for the age of AI.

4 The future of AI and DT in mathematics education assessment: key recommendations

As we have already discussed, the advent of DT and AI presents an urgent need to rethink both the purposes and the processes of learning, teaching, and assessing mathematics. Such a rethink, as the chapters collectively show, requires operating across micro and macro levels, and varied theoretical perspectives play a critical role in linking these purposes and processes. More practically, transitioning from traditional pen-and-paper assessments, which have dominated education for centuries, to digital approaches is far more than simply replicating paper-based tests in the exact same format on a screen. While the digitisation of existing formats serves a purpose (Weigand et al., 2024), we argue that it fails to capitalise on the full potential of DT and AI to transform mathematics education, and assessment in particular.

As shared in the literature (e.g. Black & Wiliam, 2010), there are three key processes in learning and teaching pertaining to assessment. These involve establishing (1) where the learners are in their learning, (2) where they are going, and (3) what needs to be done to get them there. Our analysis through the five categories identified earlier in this chapter, namely,

1. *Transitioning from traditional to digital assessment;*
2. *Assessing deeper understanding and conceptual knowledge in digital contexts;*
3. *Enhancing assessment through digital technologies;*
4. *Task design for digital and AI-based assessment; and*
5. *Human–AI complementarity,*

and the chapter's contributions under these categories encompass all of these key processes and offer insights for the digital era. For example, for assessing where learners are, technology enables assessment of knowledge and understanding as and when needed. Such information can support students in identifying what they have achieved and inform their teachers as well. This type of evaluative feedback (as mentioned in **Chapter 8**, for example), though necessary, should be accompanied with formative feedback to support students (and their teachers) about the problem-solving process, for example, rather than just the outcome. As both Hattie and Timperley (2007) and Black and Wiliam (2010) remind us, the quality and nature of any formative feedback are crucial. For establishing where learners are going and determining how to get there, the evidence presented across different chapters, especially in category (3), demonstrate the potential of DT and AI regarding the provision of immediate information to teachers and feedback to learners at scale. For example, research on computer algebra systems and automated assessment has shown both the potential and the limitations of purely automated approaches (Sangwin, 2013, and **Chapter 13**). However, we cannot ignore the danger that using DT and AI sometimes brings forward a prioritisation in speed and scale over meaningful, formative feedback. Such an argument makes us reflect upon whether the use of DT and AI is a necessity for or an enrichment of mathematical assessment. For example, **Chapter 14** questions the use of DT as they prevent students from inputting the intermediate steps in their solutions and therefore argues that some tasks are better suited for paper-based assessments, while **Chapter 12** advises moving away from 'replication and reinvention', which has characterised much of computer-aided assessment development. These arguments bring forward a first recommendation:

R1. There is a need for any evaluative (including automated) feedback offered by digital technologies and AI to be accompanied by formative feedback to students and teachers. This also implies the need for training teachers to interpret digital feedback and offer appropriate formative feedback.

Regarding AI in particular, in the last few years, the emergence of large language models (LLMs) has added another layer of complexity. Most of the field is concerned with using AI for cheating, but from a mathematics education point of view, the field should be concerned with at least the questions which we shared earlier: How can we ensure AI-based assessment focuses on process and genuine conceptual knowledge (rather than only procedural), and from a teacher perspective, how can technology be leveraged while maintaining the role of human judgement in mathematical assessment?

Even though DT and AI can support students by offering individualised feedback and even making suggestions on next steps, we do align our perspective with that of many mathematics educators that the teacher is needed in every student's learning journey. This brings us to our second key recommendation:

R2. There is a need for sustained professional development that helps teachers beyond computational and AI literacy, but to have the competency to leverage AI effectively and critically for formative and summative assessment.

Beyond these recommendations, while the potential benefits of DT are widely recognised, it is important to acknowledge that mathematics education has been slow to fully embrace their integration (e.g. Joint Mathematical Council, 2011; Wolfram, 2020). To support teachers in adopting DT in transformative ways, in the recent influential report by Royal Society UK, Crisan et al. (2023) point out to the importance of teachers developing two critical competencies: mathematical digital competency for performing mathematics with DT and mathematical digital competency for teaching mathematics using DT. In the same report, Crisan et al. (2023) make a notable recommendation that emphasises that a comprehensive integration of digital technology into mathematics education hinges on transforming the assessment process, ensuring the meaningful incorporation of DT, especially in mathematical problem-solving and reasoning. In light of these, our third key recommendation is:

R3. A comprehensive integration of digital technologies in mathematics education should particularly focus on high-stakes assessment within school mathematics, embedding digital technology actively in the assessment process.

As evidenced in this chapter and throughout this book, there is a growing interest among mathematics educators in harnessing the potential of digital tools to transform assessment practices. For example, the advent of computer algebra systems (CAS) has long raised questions about what we should assess when students can use these tools to perform computations once central to high school calculus. Should the focus shift from assessing students' mastering of methods and techniques to other mathematical skills, such as reasoning, problem-solving, or conceptual knowledge? These promising possibilities, however, remain largely absent from current mathematics assessments, both formative and summative. Why has this transformation struggled to gain traction? As we highlighted, a demand for the reimagining of the entire process of assessing mathematical learning in the digital age is needed. A critical factor lies with teachers, which leads us to our fourth recommendation:

R4. There is a pressing need to support the development of teachers' mathematical digital competency for assessing mathematics learning.

In the absence of such support, teachers are unlikely to feel both motivated and equipped to adopt digital tools into their teaching practices. We thus strongly advocate that the international mathematics education community place greater emphasis on researching and developing this competency.

To sum up, this book offers a collection of international perspectives on wide-ranging, context-specific, yet generalisable and critical arguments about computer-aided assessments, digitised assessments, and automated and adaptive feedback supported by AI. The chapters explore not only the potential of these tools but also the fundamental questions about what we assess, how we assess it, and why. By addressing these questions, we aim to inspire a shift in assessment practices that aligns with the needs of a rapidly evolving digital world.

Acknowledgements

Some of the background research and analysis that informed this chapter was supported byTransEET a HORIZON-WIDERA2021-ACCESS-03-01 project (Grant no. 101078875) funder by the EU and by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe Guarantee funding scheme. The views expressed in this chapter are those of the authors and do not reflect those of the funders.

Note

1 As also argued in Chapter 3 of this book, assessment with digital technology (or in other words, using digital technology) involves tests where students provide their solutions and answers on paper while utilising digital tools, like calculators or computers. Conversely, assessment through digital technology means that the technology is used as the platform to deliver and conduct the assessment (Drijvers, 2018).

Reference list

Aldon, G., & Trgalová, J. (2019). *Technology in mathematics teaching: Selected papers of the 13th ICTMT conference*. Springer. https://link.springer.com/book/10.1007/978-3-030-19741-4

Aspiranti, K. B., Henze, E. E., & Reynolds, J. L. (2020). Comparing paper and tablet modalities of math assessment for multiplication and addition. *School Psychology Review*, *49*(4), 453–465.

Black, P. J., & Wiliam, D. (2010). Inside the black box raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Clark-Wilson, A., Donevska-Todorova, A., Faggiano, E., Trgalová, J., & Weigand, H. G. (Eds.). (2021). *Mathematics education in the digital age: Learning, practice and theory*. Routledge. https://doi.org/10.4324/9781003137580

Crisan, C., Geraniou, E., & Hodgen, J. (2023). *Educational technologies in mathematics education. (Mathematical futures programme RFQ 5377)*. Report for the Royal Society. https://royalsociety.org/-/media/policy/projects/maths-futures/educational-technology-mathematics-education.pdf

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et évaluation en éducation*, *41*(1), 41–66. https://doi.org/10.7202/1055896ar

Drijvers, P., & Gravemeijer, K. (2005). Computer algebra as an instrument: Examples of algebraic schemes. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators: Turning a computational device into a mathematical instrument* (pp. 163–196). Springer US. https://doi.org/10.1007/0-387-23435-7_8

Drijvers, P., & Sinclair, N. (2023). The role of digital technologies in mathematics education: Purposes and perspectives. *ZDM – Mathematics Education*. https://doi.org/10.1007/s11858-023-01535-x

Entwistle, N. (2001). Styles of learning and approaches to studying in higher education. *Kybernetes*, *30*, 593–603.

Frenken, L., Libbrecht, P., Becker, B., & Greefrath, G. (2022). Dynamic geometry tasks in standardized assessment – analysis of solution processes and consequences for practice. *International Journal of Mathematical Education in Science and Technology*, 1–17.

Geraniou, E., Mavrikis, M., Jankvist, U. T., & Misfeldt, M. (2025). Generative AI in mathematics teaching: Lessons from spreadsheets through the lens of double instrumental genesis. In *Proceedings of the fourteen congress of the european society for research in mathematics education (CERME14)*. Free University of Bozen-Bolzano and ERME.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural Knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*. Lawrence Erlbaum Associates.

Jankvist, U. T., Dreyøe, J., Geraniou, E., Weigand, H.-G., & Misfeldt, M. (2021). CAS from an assessment point of view: Challenges and potentials. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalova, & H.-G. Weigand (Eds.), *Mathematics education in the digital age: Learning, practice and theory* (pp. 99–120). Routledge.

Joint Mathematical Council. (2011). *Digital technologies and mathematics education*. JMC. www.jmc.org.uk/wordpress-cms/wpcontent/uploads/2021/02/JMC_Report_Digital_Technologies_2011_Summary.pdf

Misfeldt, M., & Jankvist, U. T. (2018). Instrumental genesis and proof: Understanding the use of computer algebra systems in proofs in textbook. In L. Ball, P. Drijvers, S. Ladel, H.-S. Siller, M. Tabach, & C. Vale (Eds.), *Uses of technology in K-12 mathematics education: Tools, topics and trends. ICME-13 monographs* (pp. 375–385). Springer. https://doi.org/10.1007/978-3-319-76575-4_22

Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment* (pp. 92–98). European Commission Joint Research Centre.

Sangwin, C. (2013). *Computer aided assessment of mathematics*. Oxford University Press.

Shute, V. J., & Becker, B. J. (Eds.). (2010). *Innovative assessment for the 21st century*. Springer.

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*, *77*(1), 20–26.

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (Vol. 27, pp. 721–751). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal*, *42*(3), 454–476. https://doi.org/10.1002/berj.3215

Weigand, H. G., Trgalova, J., & Tabach, M. (2024). Mathematics teaching, learning, and assessment in the digital age. *ZDM Mathematics Education*, *56*, 525–541. https://doi.org/10.1007/s11858-024-01612-9

Wolfram, C. (2020). *The math(s) fix: An education blueprint for the AI age*. Wolfram Media, Inc.

2

THE TRANSFER OF ASSESSMENT FROM PEN AND PAPER TO E-ASSESSMENT

Two case studies from one theoretical perspective

Paola Iannone, Alice Lemmo, and George Kinnear

1 Introduction

In this chapter, we focus on how and why students' engagement and outcomes may differ when solving tasks in different modalities: In particular, we compare paper-based and e-assessment tasks. These differences can have important consequences, both in terms of how knowledge is shaped by the tools the students use when working on the task and in terms of the validity of the assessment. We propose the instrumental approach (Trouche, 2005) as a tool to analyse these differences. To this end, we first introduce one classification of task transfer, we then present briefly the main aspects of the instrumental approach (Trouche, 2005), and finally we use this framework for the analysis of two case studies of task transfer. We conclude with some remarks about the potential of using this framework to analyse what happens when we transfer tasks from paper-based to e-assessment.

The motivation to focus on task transfer is that the integration of digital technologies into teaching and learning has significantly expanded the possibilities of task design, among other things, in mathematics education (Hoyles & Lagrange, 2010). In particular, e-assessment[1] is now widespread at all educational levels. As Lowrie and Logan (2015) noted, the benefits of e-assessment include cost reduction, faster feedback, and the ability to customise tests to individual needs. As recent educational research is increasingly focused on teaching, learning, and assessing digital competencies (Voogt & Pareja Roblin, 2012), Hoyles et al. (2010) introduced the concept of techno-mathematical literacies, which refers to mathematical activities involving technological tools. In this context, e-assessment appears to be a promising method for assessing such skills alongside mathematical skills (Gane et al., 2018).

DOI: 10.4324/9781003533764-2

However, the issue of interpretation of outcomes of e-assessment may pose difficulties. When the scores obtained from a test in e-assessment mode are expected to be interpreted in the same way as the scores from an assessment in paper-based mode, it becomes essential to gather appropriate evidence to support the reliability and validity of e-assessment outcomes. The research evidence (Lynch, 2022) suggests the need for a paradigm shift, since what is known about outcomes of tasks in paper-based mode cannot be transposed to the digital environment. Issues may arise when transferring a task from paper-based mode to e-assessment mode in terms of the skills assessed. It is important to understand the nuances of these changes, which are not only due to familiarity with the devices, the type of task, or the age of the students involved. The aim of this chapter is to describe and better understand this issue through a theoretical framework devised to analyse the interactions of the humans engaged in solving a task requiring the use of artefacts (not necessarily digital). In what follows we will first discuss ways of transferring tasks from paper-based to e-assessment, and then we will introduce the theoretical framework we have adopted for the analysis of the two case studies in this chapter.

2 Transferring tasks from paper-based to e-tasks

Much of the literature concerning e-assessment focuses on validity. In a naïve sense, *validity of assessment* can be defined as the degree to which the assessment measures the construct it was designed to measure. For example, taking a test on a computer requires some degree of computer skills, and the proficiency with these skills, or lack thereof, could potentially be captured in the outcome of the assessment. If the tool which administers the assessment (e.g. a digital tool) introduces additional constraints or requires additional capabilities to succeed than that of the paper-based assessment, validity of that assessment may be jeopardised as the e-assessment may not be assessing (only) the construct it is meant to be assessing. This situation is called construct-irrelevant variance (Messick, 1995). Lemmo's (2021) study shows that construct-irrelevant variance can affect assessment outcomes. The specific type of construct-irrelevant variance introduced by the mode of assessment administration is often referred to as a mode effect. *Mode effect*, in the broadest sense, is 'any difference found in test performance that is attributed to the mode of administration' (Way et al., 2015, p. 263).

Because of the importance of mode effect, we turn now to the issue of task transfer and consider examples of task transfer when a (mathematical) task that has been designed to be administered in the traditional way as paper-based assessment is transferred to e-assessment.

While there are various frameworks related to task transfer (e.g. Romrell et al., 2014), we focus on one that we consider most relevant for e-assessment: the one by Ripley (2009). In this framework, there are two strategies identified for task transfer:

- **The migratory strategy** is defined as taking a paper-based assessment task and putting it in a digital environment while living the task 'qualitatively unchanged'. (p. 94). This strategy does not seek to change the curriculum, teaching, or learning.
- **The transformational strategy** supports the integral use of digital technology (such as applets or dynamic geometry systems) in assessment. This assessment is 'designed to influence (or minimally to reflect) innovation in curriculum design and learning' (p 94). However, Ripley (2009) noted that the potential for educational innovations linked to the use of e-assessment was – at the time of writing – not yet realised.

Ripley's (2009) classification is to be understood as a continuum where both the mode of assessment (paper-based or e-assessment) and the level of innovation that the assessment proposes come into play.

Example of migratory strategy. A question that could be asked as paper-based assessment is:

Given $f(x) = \frac{1}{x+5} - 5$, find the inverse function, $f^{-1}(x)$.

This could instead be asked in the form of e-assessment; for instance, Figure 2.1 shows an implementation of this task using STACK (Sangwin, 2013).

The example in Figure 2.1 demonstrates the migratory strategy in that a paper-based assessment item was migrated to an online interface.

Example of transformational strategy. In Figure 2.2, we have an example that could be considered a transformation of a paper-based assessment to e-assessment. In the paper-based assessment, students are asked the question:

Find a function f: [0,1] → [0,1] with image [0,½].

For the e-assessment version, students are asked to find a function with the same properties using an applet that shows a grid with four movable points (see Figure 2.2). We consider this a transformation of the task, since the applet restricts the space of functions that can be used.

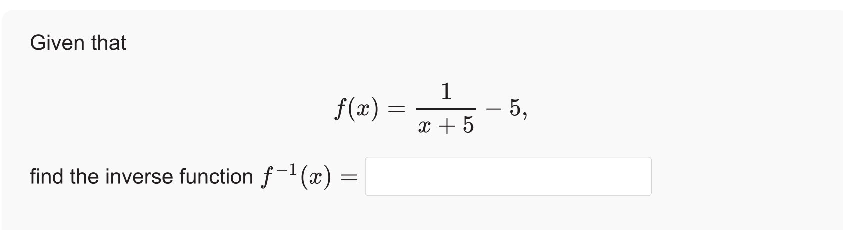


FIGURE 2.1 Example of the STACK interface in the case of the migration of a task – the answer is written in the white box after the equals sign in the online interface.

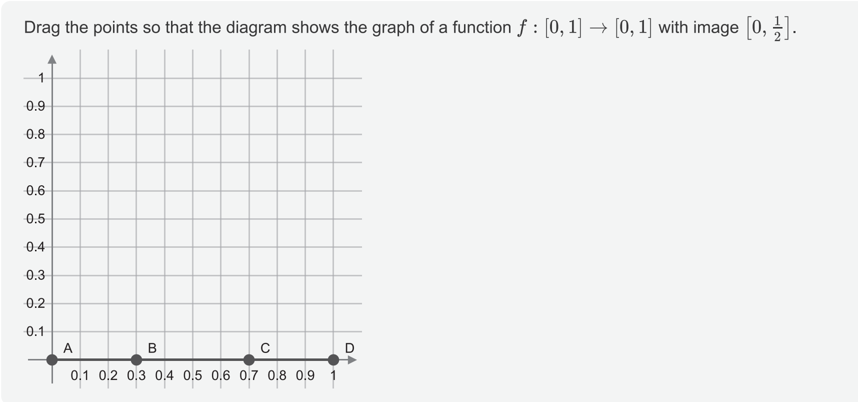


FIGURE 2.2 An e-assessment task implemented using STACK in the case of transformation of a task. The applet allows students to move the 4 points on the grid to find the desired function.

The education literature (e.g. Ripley, 2009; Bennett, 2015) suggests that different levels of innovation in assessment modes may lead to innovation and change in students' educational experience. However, the question concerning validity remains: Does the translation introduce construct-irrelevant variance? To address this question, we consider two case studies where we focus on how students interact with paper-based and e-assessment tasks. These two case studies aim to be an exemplification of the kind of information a specific theoretical framework could provide in terms of skills required by the assessment with the change of mode, but they are not meant to be exhaustive of all that could happen in the transfer. We also recognise that the characteristics of the e-assessment tool will have an impact on the capabilities (or change of capabilities) assessed.

3 The instrumental approach

The focus of the chapter is on the difference in students' interactions with two distinct tools in two case studies and the way in which the interaction with the tool shapes how students use or manifest their knowledge and skills. If we find that the interaction that students have with the e-assessment tool in solving the task is substantially different from that which occurs when students solve the same task proposed in a paper-based mode, then issues of validity and fairness may arise. To analyse such interactions, we use the instrumental approach (Trouche, 2005). This framework has been fruitfully used to investigate the interaction of students and graphic calculators in the early

2000s (e.g. Guin et al., 2005) and more recently to investigate the interaction between students and programming tools for learning mathematics (Gueudet et al., 2022; Thoma & Iannone, 2023). This theoretical framework draws on activity theory (based on Vygotsky's 1978 work) and the impact on knowledge creation of the interaction between human and tool. In this case, an artefact (e.g. the e-assessment interface, or the pencil and paper) becomes an instrument (the artefacts together with the ways of working with the artefact that a human develops to fulfil a goal) when schemes are established that regulate such interactions. The schemes are defined as adaptable organisation models (ways of working) that account for the interaction between user and tool in each given situation (Vergnaud, 2009). The schemes are composed of four aspects:

1. **Intentional aspect.** The goal(s) of the activity. In our work, we assume that the goal of the activity is to solve one or more assessment tasks (in e-assessment or paper-based mode).
2. **Generative aspect.** The rules-of-action, which are 'the sequences of actions, information gathering, and controls' (Vergnaud, 2009, p. 88) that a human performs while interacting with a tool. In our case, these are actions that a student performs while working with the task. A rule-of-action that a student performs while solving a mathematical task on a computer-aided system could be to constantly refer to the definition of the mathematical objects that are involved in the task before starting to solve the given task.
3. **Epistemic aspect.** The operational invariants, which are concepts-in-action and theorems-in-action. In our case, theorems-in-action are the rules that the student holds true and that guide their actions, while concepts-in-action are the concepts that the student holds relevant to the current situation. We note here that the word *theorem* does not imply only mathematics theorems but also other rules that a student may hold true and which may, for example, relate to rules about the use of the tool. A theorem-in-action that a student may hold while solving a mathematical task concerning finding an injective function that satisfies certain properties on a computer-aided system is that 'for such function each element in the domain gets mapped only in one element of the codomain'. Among the related concepts-in-action here would be that of 'function', 'domain', and 'codomain'.
4. **Computational aspect.** The possibilities of inferences from the actions and operational invariants (theorems-in-action and concepts-in-action).

In adopting this framework, we operationalised the elements described earlier by creating codes for the rules-of-action, theorems-in-action, and concepts-in-action that we could discern in the data and drawing inferences with respect to the different instrumentations we could find.

The process of instrumental genesis concerns the emergence of such schemes. Finally, *instrumentation* is defined in Trouche (2004) as the

> process by which the artifact prints its mark on the subject, i.e., allows him/her to develop an activity within some boundaries (the constraints of the artifact).
>
> *(p. 290)*

In this chapter, we aim to document the process of instrumental genesis as it unfolds when a task is initially presented using traditional tools, such as pen and paper, and subsequently adapted into an e-assessment format—particularly in cases where this shift constitutes a transformation (Ripley, 2009).

4 Case study 1: hit the balloon task – find the midpoint of a segment

The first case study examines the transformation of a task given to first year middle school students (grade 6 in the Italian system). The pupils' teachers stated that the task of finding the midpoint of a segment without an explicit metric is new for students, in both pen-and-paper and e-assessment modes. This case study explores the impact of the task transformation on the students' approaches and outcomes. Detailed results of this research are discussed in Lemmo (2020, 2023), where the problem-solving framework proposed by Schoenfeld (1985) was adopted for the analysis of the data. The results showed similarities in the strategies adopted, but differences in the phases (episodes) of the solving process, depending on whether the students worked on the task using pen and paper or e-assessment. We show here how the instrumental approach can offer insight into why the differences occur in the two groups and bring to the fore the difference in task approaches. We present the analysis of a task inspired by an applet developed by the Freudenthal Institute Research Group in Mathematics Education: 'Hit the balloon' task[2] (Figure 2.3).

The goal of the task is to find an estimate of the position of a numeric value on a number line in which the metric is not made explicit but only the value of the two extreme points on the line. Students were asked to work in pairs: Some pairs performed the task in paper-based mode, while others used the applet. In the following, we discuss for each of the environments (paper-based and e-assessment) one of the related schemes activated by the students.

The 'unit of measure' scheme. In the paper-based mode, most students use the ruler to identify a metric on the line. In this case, the students add a tool to their work: the ruler. In the example presented in Table 2.1, students use the ruler with the aim of identifying a metric of 10. Indeed, they search for the length corresponding to the value 10, after determining the position of

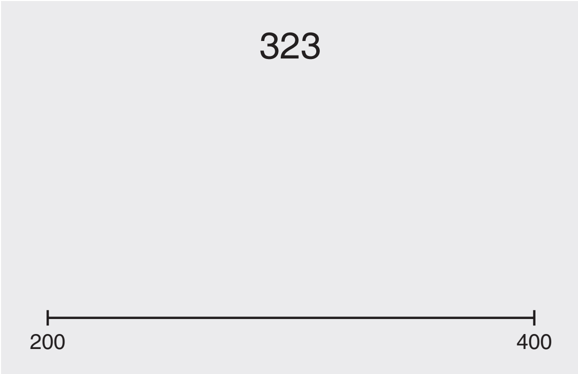


FIGURE 2.3 Task for grade 6 students: In the paper-based version, the assignment is: ‘Draw a cross in the point on the line that corresponds to the number you see on the top.’ In the digital version, the assignment is: ‘Move the cross in the point on the line that corresponds to the number you see on the top.’

TABLE 2.1 Students N and L dealing with task 1 in paper-based mode, using also ruler for this task

Transcript – Task 1
Task 1: Identify the position of 335 in a range between 200 and 400.
Students (N and L) identify the value of the middle point as 300; then they take the ruler.
N: Uh! There is a ruler!
L: Just do something. There’s also the pencil. Measure.
N: Then it is the exact half; the segment 12 cm long.
Then the exact half is 6 cm because it [segment] is 12 cm, more or less.
L: Oh, I don’t like ‘more or less’. It is 12.4 cm long!
N: Okay, well, never mind. You want to be perfect.
To identify the 35 position, we have to find . . .
L: Ten, ten, ten.
N: That’s right, we should find out how much 10 values on the line.

the midpoint (300) in the segment [200, 400]. In the transcript, we notice the declared intention to be precise and not to estimate the position.

Most of the students who solved the paper-based task (the intentional aspect is to find such a solution) used the rulers to identify a precise metric, and these actions often lead to difficulties in identifying the correct position of the required point. Here, one of the rules-of-action detected is 'We need the ruler to measure precisely' (generative aspect), while the corresponding theorem-in-action is 'Only with a ruler can I find an exact metric to solve the problem', and the related concept-in-action is that of precise measurement (epistemic aspect). This scheme, therefore, involves rules-of-action such as 'I need to measure precisely' and 'I need to use the ruler to identify a scale on the given segment'. The relevant concept-in-action is that of precise measurement, and related theorems-in-action include 'The midpoint of the segment corresponds to the midpoint between the values of the two extremes'. After several attempts to find the metric (computational aspect), most students then decided to switch to the bisection method. It was only after the failure to complete the task by deploying the 'unit of measure' scheme that some pairs reverted to an estimate of the midpoint by bisection.

The 'bisection' scheme. In the e-assessment mode, students immediately approached the task by finding the midpoint and then dividing the line further in halves until an appropriate approximation is reached. In this case, the time spent on the task was very short for all students, and the accuracy of the results was satisfactory. In the example presented in Table 2.2, the students identify the midpoint (corresponding to 300) and then 3/4 of the line (corresponding to 350). Finally, they find the position of 325 and place the cross just to its right.

In this scheme, rules-of-action (generative aspect of the scheme) include the estimate of relevant points on the segment ('I identify points of interest on the segment (quarters, thirds, and so on)'); they all relate to estimates and not correct measurements – despite a ruler being available to those students as well, as can be seen in the image in Table 2.2. Theorems-in-action also refer to estimates ('There is a correspondence between the length of a portion of a segment and a certain range of units') linking the interaction to estimating and bisecting (epistemic aspect of the scheme).

Eventually, in both assessment modalities, students identified the solution by successive approximations (computational aspect – the solution of the goal). In the case of the paper-based mode, this process was slowed down by using the ruler, which induces the students to define a metric on the line and engage in difficult divisions in the attempt of being precise. In the case of the e-assessment, no student attempted to identify a metric: The pairs proceeded by bisecting segments, despite a ruler being available. In many studies, the ability to deal with estimates on the number line is correlated with mathematical competence (e.g. Schneider et al., 2018). The paper-based task seems to

TABLE 2.2 Students A and F dealing with task 1 in e-assessment mode

Transcript – Task 1
Task 1: Identify the position of 335 in a range between 200 and 400.
Students (F and A) identify the value of the middle point as 300, and they proceed by further dividing the line in halves.
F: So 300 is more or less . . . then . . .
A: Yes, in the middle there is 300.
F: . . . a little further ahead.
A: So here is 300. And here is the 350.
F: Wait, more or less half [*referring to the half between 1/2 and 3/4 of the line*].
A: Eh, here is the 350, and here should be the 335.
F: 335, yes.


limit or initially inhibit this estimation process, while the e-assessment mode seems to favour it. If we were to think of this task in the context of timed assessment, the mode effect (Way et al., 2015) brought by the use of the paper-based task would disadvantage the students.

5 Case study 2: example generation on STACK

The second case study concerns the transformation of four example-generation tasks administered to first year students in mathematics at a university in the UK, and the impact that such transformation had on students' approaches to and the outcome of the tasks. We note that in this university, students are familiar with STACK, which is used both for formative and for summative assessment. The results of the larger project where this case study comes from have been elaborated in Kinnear, Iannone, et al. (2024), where we analyse the students' engagement with the tasks via an investigation of the strategies (as in Antonini, 2011) they employ to produce examples. In this chapter, we re-analyse some of the data using the instrumentation framework (Trouche, 2005) to map possible reasons for the differences observed. The study involved two groups of students on the same first year course at the same university who attended task-based interviews. One group was asked to solve the four tasks in Figure 2.4 in paper-based mode, and the second was asked to solve the same tasks using a STACK interface,[3] where they could move four points on a grid (see Figure 2.2).

In a larger study adopting the same materials, we observed that the students who engaged with the tasks in the paper-based version did significantly

In each case, draw the graph of a function with the given properties, and label important points:

- $f_1 : [0, 1] \to [0, 1]$ has image $[0, \frac{1}{2}]$.
- $f_2 : [0, 1] \to [0, 1]$ has image $[0, \frac{1}{2}]$ and is not injective.
- $f_3 : [0, 1] \to [0, 1]$ is surjective and not injective.
- $f_4 : [0, 1] \to [0, 1]$ is injective, not surjective, and passes through $(0.2, 0.8)$ and $(0.5, 0.5)$.

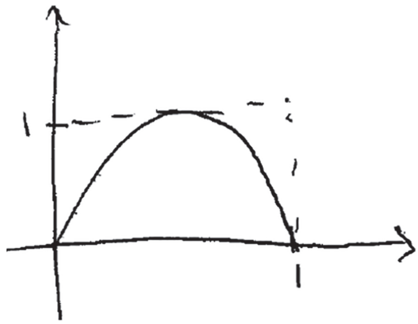
FIGURE 2.4 The text of the tasks administered to the two groups of students.

better than those who engaged with the tasks on the STACK interface (Kinnear, Iannone, et al., 2024). Here we propose a way of analysing the interview data to detail the instrumental genesis that occurs in the two groups to offer a possible explanation of why such a difference was observed. The goal of the activity was to produce an example of a real valued function with the desired characteristics (intentional aspect of the framework). In what follows we discuss two distinct instrumentations and related schemes: one for the paper-based task, and one for the e-assessment task.

The 'analysing and retrieving' scheme. The analysis of the interviews with students who solved the tasks in paper-based mode reveals that for those students the prevalent scheme to solve the given tasks consisted in analysing the definition of the concepts relevant to the task and then retrieving examples from their example spaces. This is not surprising, as example-generation sequences have been designed to help students enrich their example space (Watson & Mason, 2005), and tasks of example generation have been used in research to explore example spaces. An example that illustrates this strategy is in Table 2.3.

In Table 2.3 we notice that the student focuses their actions on analysing what the task requires from them, reading the definitions involved from the lecture notes provided. The main sequence of actions concerns the analysis of the definition relevant to the task ('I need to refer to the definition of *surjective* and *injective'* – generative aspect of the framework). Theorems-in-action related to this scheme are 'A function needs to go through every value in its codomain' and 'For a function not to be injective, there needs to exist a point in the codomain that is the image of two points in the domain'. Concepts-in-action relevant to the scheme are that of domain, codomain, injective, and surjective (epistemic aspects). The final action is retrieving an example the student is already familiar with which fulfils the requirements of the task (computational aspects – possibility of inference). This is signified by the line 'I could just have another parabola'. This scheme was common for students who solved the task in paper-based mode, and we argue that the tools (pen and paper in this case) facilitate the retrieval of examples from the example space. Success or otherwise in the task therefore depended also on the richness of the example space.

TABLE 2.3 Student P7 attempting the task in paper-based mode

Transcript annotated with student's example	
Task 3: Draw the graph of a function $f_3 : [0,1] \to [0,1]$ that is surjective and not injective.	
P7: It's surjective and not injective. So it has to go through every point in the codomain. Which actually ends at 1 this time, because we haven't got a smaller image, because the image would have to equal . . . It goes through every *y* point here . . . em, goes up to 1 [*annotating the x-axis*]. So it goes through every *y* point we need. Point where . . . we need a point again where this is true [*annotating the x-axis*]. So we have a *y* value that is repeated at two different *x* values . . . but I could just have a . . . [*draws curve*] another parabola. [*The student had also used a parabola for Task 2.*]	

The 'point-by-point construction' scheme. Here we see how the instrument (the STACK interface in this case) has influenced the solution of the tasks. In the extract in Table 2.4, we notice how the student constructs the example point by point for the four points given in the grid. We argue that the student is not retrieving an existing example from their example space – but they (try to) construct new ones given the potential and constraints of the tool.

The first action performed is that of referring to the previous example, analysing it for its properties and then changing it by moving the points B and C to make a function which is not injective ('I can use and modify previous examples to solve a new task' – generative aspect). As for the previous case, the concepts-in-action relevant to the task are those of function, injective, and image, whereas the theorems-in-action include the horizontal test for injective functions (epistemic aspect).

We notice from the analysis of the whole body of data that students in the e-assessment mode did not appear to retrieve an example from their example space and adapt it to the e-assessment mode. In particular, examples based on parabolas were common in the paper-based task (as in Table 2.3), and it did not seem that students working in the e-assessment mode were retrieving these examples then fitting them to the STACK interface (e.g. by drawing a cusp). Instead, we observed a different instrumentation, where students constructed the required example point by point, as in the 'point-by-point'

TABLE 2.4 Student S1 solving the task in the STACK interface

Transcript annotated with student's example

Task 2: Draw the graph of a function $f_2 : [0,1] \to [0,1]$ that has image [0, ½] and is not injective

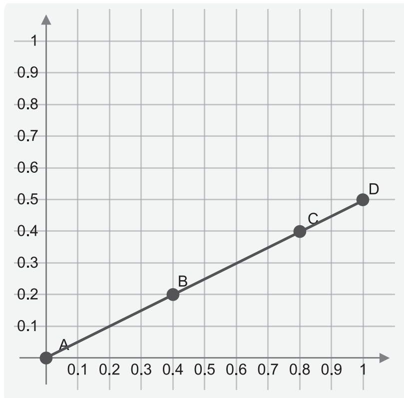
S1 reads the question, then talks through the definition of 'injective'.

S1: I'm just thinking about my last one [*previous task*], and whether that was injective or not, because that was just a straight line. . . . I don't think there is anything . . . I think my last one was injective. I was thinking, in case the last one was, I could draw the same function. . . .

I: Yes, we agree that the last one was injective. So what would you have to do this time?

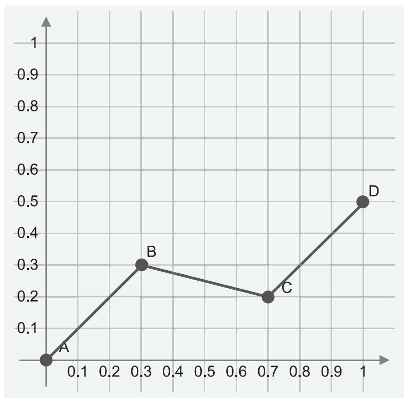
S1: I'm just going to put in my last one, just to see if that helps me think about it

[*student moves the 4 points to make a straight line*].



S: The same function isn't going to work for this. It's not injective, then it needs to . . . there needs to be a point on the graph where you can draw a horizontal line across it and it'll hit the graph twice. . . . It needs to change direction at some point, basically [*moves points B and C*].

S: I would want it to look something like this.



Note: Like the students in the paper-based tasks, these students had also been given the lecture notes from the course so that they could refer to them. In the dialogue, *I* is the interviewer.

scheme presented earlier. The students using the STACK interface were restricted to working with the example space of piecewise linear functions; given that students find these functions difficult to work with (Hohensee, 2006), we hypothesise that the restrictions imposed by the interface made it more difficult for students to retrieve a suitable example (Kinnear, Iannone, et al., 2024). This is an example of mode effect: A task that was designed to assess the richness of students' example spaces is now assessing the example space restricted to piecewise linear functions, a task which appeared to be more difficult for students.

6 Conclusions

In this chapter we have employed the instrumental approach to investigate the transfer of tasks from paper-based to e-assessment. To do so, we have analysed data from two case studies at different educational levels. In both case studies, we found a mode effect (Way et al., 2015) which potentially had an impact on the validity of the assessment. In the first, the paper-based task slowed down students who attempted (unsuccessfully) an exact measurement of the given segment, since the students developed substantially different schemes during the solution of the task in the two modalities. The students who solved the task in e-assessment employed a bisection scheme, which was more efficient and, therefore, would have taken less time in the context of a timed assessment. In this case, the paper-based mode seems to be more useful than the e-assessment mode for assessing students' ability to construct metrics, but not in making estimates. In the second case study, the students who solved the paper-based example-generation task were able to access a wider example space to find the required functions than did the students who solved the task in e-assessment, and the schemes that were developed show this difference. Here the mode effect concerned the difference in construction of the example required: The e-assessment seemed to disadvantage students, as they appeared to construct examples point by point rather than retrieving them from their example space. In both cases, the transfer to e-assessment has the potential to impact the validity of the task. However, we note that the mode effect can facilitate targeted tasks: For example, in the second case study, the e-assessment would have helped students become more familiar with piecewise linear functions, which would not have been the case in the paper-based assessment.

We are aware that the degree to which a measure remains consistent across different test media depends on the specific measure, the participants involved, the software and hardware used for testing, and so on (Lenhard et al., 2017), but the theoretical lens adopted in this study contributes to better understand such differences in terms of generative, epistemic, and computational aspects. In the first case study, the instrumental approach allows us to take the analysis beyond the generative aspects, hence the strategies used in solving the task (in the example, the bisection method), by looking at rules-of-action and concepts/theorems-in-action (epistemic aspects). A subsequent study may help us to further investigate this evidence and, in particular, the relationship between rules-of-action and concepts/theorems-in-action. In the paper-based mode, it appears that concepts/theorems-in-action guide the rules-of-action, such as the need to be precise and to measure. On the other hand, in the e-assessment mode, the different rules-of-action imposed by the environment seem to lead students to choose different concepts/theorems-in-action, such as the correspondence between length and value on the line. This aspect confirms previous studies (Lemmo, 2023) in which the mathematical resources activated by students in the two modes are different and therefore induce them to adopt different approaches and strategies. Similarly, for the second case study, the epistemic aspects appear to affect generative and computational aspects: In the paper-based mode, the theorems/concepts-in-action related to the concepts relevant to the task (injectivity, surjectivity, function, etc.) seemed to influence the rules-of-action of retrieving a suitable example, while in the e-assessment mode, the actions allowed by the e-assessment (dragging the given points) drove actions related to point-wise construction of the function required. Again, the mathematical resources activated by the two modes are different, and this may account for the difference in outcomes observed by Kinnear, Iannone, et al. (2024).

This chapter also offers some research directions for future work: e-assessment allows to test skills that are at times different from those tested by the paper-based assessment. In both of our case studies, we saw the emergence of very different schemes when students engaged with paper-based and e-assessment tasks. Awareness of the type of support provided by technologies can and should give indications as to what is the most suitable environment in which to develop, observe, and assess certain competences, knowledge, and mathematical skills. The points raised in this chapter have the potential to significantly influence educational policy and assessment research, particularly concerning task design. When interpreting scores from e-assessment mode, especially from large-scale national assessment, in the same way as those from assessment in pen-and-paper mode, it is crucial to collect appropriate evidence to ensure the reliability and validity of e-assessment mode's scores. A multiplicity of theoretical frameworks can enable the construction of a broader and more specific awareness. However, such awareness can only be created by careful consideration of the interaction between students and tasks, as our case studies show.

Notes

1 We use here the term *e-assessment* in the same meaning as Kinnear, Jones, et al. (2024) to indicate any assessment which involves the use of computer technology. We use the term *paper-based* assessment for assessment that is administered in the traditional pen-and-paper mode.
2 www.fisme.science.uu.nl/toepassingen/28230/.
3 The tasks are available at https://osf.io/a6y2g/.

Reference list

Antonini, S. (2011). Generating examples: Focus on processes. *ZDM*, *43*(2), 205–217. https://doi.org/10.1007/S11858-011-0317-6

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education, 53*(2), 176–187. https://doi.org/10.1111/ejed.12269

Gueudet, G., Buteau, C., Muller, E., Mgombelo, J., Sacristán, A. I., & Rodriguez, M. S. (2022). Development and evolution of instrumented schemes: A case study of learning programming for mathematical investigations. *Educational Studies in Mathematics, 110*(2), 353–377. https://doi.org/10.1007/s10649-021-10133-1

Guin, D., Ruthven, K., & Trouche, L. (2005). *The didactical challenge of symbolic calculators: Turning a computational device into a mathematical instrument* (Vol. 36). Springer Science & Business Media.

Hohensee, C. (2006). Students' thinking about domains of piecewise functions. In S. Alatorre, J. L. Cortina, M. Sáiz, & A. Méndez (Eds.), *Proceedings of the 28th North American chapter of the international group for the psychology of mathematics education*. Universidad Pedagógica Nacional.

Hoyles, C., & Lagrange, J. B. (2010). *Mathematics education and technology: Rethinking the terrain*. Springer.

Hoyles, C., Noss, R., Kent, P., & Bakker, A. (2010). *Improving mathematics at work: The need for techno-mathematical literacies*. Routledge.

Kinnear, G., Iannone, P., & Davies, B. (2024). Student approaches to generating mathematical examples: Comparing e-assessment and paper-based tasks. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-024-10361-1

Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., Foster, C., Heck, A., Henderson, K., Hunt, T., Iannone, P., Kontorovich, I., Larson, N., Lowe, T., Meyer, J. C., O'Shea, A., Rowlett, P., Sikurajapathi, I., & Wong, T. (2024). A collaboratively-derived research agenda for e-assessment in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education, 10*(1), 201–231. https://doi.org/10.1007/s40753-022-00189-6

Lemmo, A. (2020). From paper-and-pencil to computer-based assessment: An example of qualitative comparative analysis. In A. Donevska-Todorova, E. Faggiano, J. Trgalova, Z. Lavicza, R. Weinhandl, A. Clark-Wilson, & H. Weigand (Eds.), *Proceedings of the 10th ERME topic conference MEDA, Linz, Ausria* (p. 385). Johannes Kepler University.

Lemmo, A. (2021). A tool for comparing mathematics tasks from paper-based and digital environments. *International Journal of Science and Mathematics Education, 19*(8), 1655–1675. https://doi.org/10.1007/s10763-020-10119-0

Lemmo, A. (2023). Tasks in paper and digital environments: An exploratory qualitative study. *International Journal of Mathematical Education in Science and Technology, 56*(4), 667–687. https://doi.org/10.1080/0020739X.2023.2268613

Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of screen versus print reading comprehension depends on task complexity and proficiency. *Discourse Processes, 54*(5–6), 427–445. https://doi.org/10.1080/0163853X.2017.1319653

Lowrie, T., & Logan, T. (2015). The role of test-mode effect: Implications for assessment practices and item design. In pursuit of quality mathematics for all. In M. Berger, K. Brodie, V. Frith, & K. le Roux (Eds.), *Proceedings of the 7th ICMI-East Asia regional conference on mathematics education* (pp. 649–656). Philippine Council of Mathematics Teacher Educators (MATHTED).

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research & Evaluation, 27*, 22. https://doi.org/10.7275/pare.1317

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 92–98). Office for Official Publications of the European Communities.

Romrell, D., Kidder, L., & Wood, E. (2014). The SAMR model as a framework for evaluating mLearning. *Online Learning Journal*, *18*(2). https://doi.org/10.24059/olj.v18i2.435

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford University Press.

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, *89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Thoma, A., & Iannone, P. (2023). Natural number game: Students' activity using an interactive theorem prover. In P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, & E. Kónya (Eds.), *Proceedings of the thirteenth congress of the european society for research in mathematics education (CERME13)* (pp. 2543–2550). Alfréd Rényi Institute of Mathematics and ERME.

Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning*, *9*, 281–307. https://doi.org/10.1007/s10758-004-3468-5

Trouche, L. (2005). An instrumental approach to mathematics learning in symbolic calculator environments. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators: Turning a computational device into a mathematical instrument* (pp. 137–162). Springer. https://doi.org/10.1007/0-387-23435-7_7

Vergnaud, G. (2009). The theory of conceptual fields. *Human Development*, *52*(2), 83–94. https://doi.org/10.1159/000202727

Voogt, J., & Pareja Roblin, N. N. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Vygotsky, L. (1978). *Mind in society*. Harvard University Press.

Watson, A., & Mason, J. (2005). *Mathematics as a constructive activity*. Routledge.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In *Technology and testing* (pp. 260–284). Routledge.

3

DIGITAL TECHNOLOGY AND ASSESSMENT VALIDITY

Exploring utilisation schemes for a basic spreadsheet tool used in PISA 2022

Mattias Winnberg, Samuel Sollerman, and Hendrik Van Steenbrugge

1 Introduction

Assessment in mathematics education is characterised by a well-accepted duality between assessment *with* digital technology and assessment *through* digital technology (Drijvers, 2018; Stacey & Wiliam, 2013). Assessment with digital technology refers to tests where students present their solutions and answers using pen and paper, with access to digital technology (e.g. calculators or computers). In contrast, in assessment through digital technology, the technology serves as a testing environment to deliver and administer the assessment (Drijvers, 2018). This chapter will focus on assessment through digital technology, especially when digital tools are integrated into large-scale assessments.

One of the key drivers behind advancing digital assessment in mathematics education is the need for more authentic assessments (Stacey & Wiliam, 2013). Such assessments prioritise students' competencies in performing 'real-life' mathematics, particularly involving digital technologies to address mathematical problems, rather than relying heavily on multiple-choice questions. This shift is evident in large-scale assessments, such as PISA 2022, where digitalisation has created opportunities for interactive software, for example, spreadsheets. These tools enable the design of dynamic and interactive test items that would presumably be impossible to present using traditional paper-based methods (Weigand et al., 2024). Proponents of this approach (e.g. Stacey & Wiliam, 2013) argue that authentic assessments can emphasise mathematical concepts by reducing the 'burden' of routine calculations, allowing students to focus on higher-order skills, such as problem-solving, reasoning, modelling, and argumentation. However, others (e.g. Jankvist &

DOI: 10.4324/9781003533764-3

Misfeldt, 2015) caution that overemphasising technical aspects may weaken the connection between procedural fluency and conceptual understanding. This concern arises when technology-based activities prioritise procedural mastery at the expense of mathematical processes and conceptual understanding. In contrast, some scholars (e.g. Drijvers & Gravemeijer, 2005) argue that separating technical skills from conceptual understanding is neither rational nor desirable when using digital tools. This raises the critical question of whether it is possible – or even reasonable – to treat technical skills and conceptual understanding as distinct entities in the context of digital assessment through digital technology.

In this chapter, we build on the reasoning of Drijvers and Gravemeijer (2005), who emphasise the intertwined and co-evolving nature of techniques and conceptual understanding when digital tools are integrated into mathematical problem-solving activities. One of their key arguments is that attempting to separate techniques from conceptual understanding – and outsourcing techniques to a digital tool to focus on conceptual aspects in assessment – may be a rather naive approach. Along this line, we examine this argument by investigating an assessment (exemplified by situations from the PISA 2022 assessment) where students receive technical training to use a digital tool before the assessment. We focus on a specifically chosen spreadsheet tool, because of its seemingly straightforward handling requirements applied in one of the most difficult items of PISA 2022. Our analysis emphasises the interplay between technical and conceptual aspects of digital tool use (see Section 4). It examines what is being assessed, particularly regarding the validity of this assessment in measuring mathematical competence (see Section 3). Since questions about what is being assessed are inherently linked to validity, this chapter situates itself within the broader literature on the validity of digital assessment, while also addressing their opportunities and challenges.

1.1 Aim and research questions

This chapter aims to investigate what is assessed when digital tools are integrated into large-scale assessments and how this influences the validity of inferences made from test scores. The chapter will be guided by the following research questions:

1. What technical and conceptual aspects must students develop to use a spreadsheet tool productively in the PISA assessment?
2. What technical and conceptual aspects have students been able to develop through the opportunities provided in a training section?
3. In light of (1) and (2), how will this potentially influence assessment validity?

2 Digital assessment – opportunities and challenges

The increasing integration of technology in today's society has led to the greater incorporation of digital tools in educational settings and, therefore, mathematics assessment. Digital assessment in mathematics is driven by the possibilities they offer in terms of item selection (e.g. adaptive testing), presentation methods (e.g. video stimuli), and response formats, which allow students to create, manipulate, and analyse graphs and geometrical objects (Drijvers, 2018; Stacey & Wiliam, 2013). These advancements also allow test developers to design richer and more interactive test items (Drijvers, 2018; Weigand et al., 2024; Yerushalmy et al., 2017), while automated scoring can reduce the workload for test graders (Drijvers, 2018). Generally, digital assessments and the integration of digital technology are often argued to broaden what can be measured (Beller, 2013; Bennett, 2002). These advancements allow assessments to better align with learning activities in digital contexts.

However, several challenges with digital assessments are well-recognised in mathematics education and must be carefully considered. These challenges include limitations in the assessment environment, such as constraints of automated scoring, which may lead to non-authentic items, testing of less advanced mathematical aspects (Pepin et al., 2017), and the fact that available software often lacks the capability for students to express mathematical ideas using a combination of algebraic and graphic representations along with natural language (Drijvers, 2018). It may also pose challenges if the software creates too great a distance from students' conventional learning environments. This instrumental distance (Haspekian, 2005) then functions as a constraint for students in the assessment environment. Additionally, there is a possibility that mathematical problems may become more complex due to the greater variety of representations in digital formats and an increased emphasis on conceptual knowledge, which is generally considered more challenging for students (Weigand et al., 2024). Despite efforts to use digital technology in assessments to measure higher-order skills, many of these assessments remain relatively inflexible, with a high proportion of multiple-choice questions. Although such assessments are valued for their stability and reliability, they may not be adequate indicators of real-life performance (Hoogland & Tout, 2018).

In conclusion, digital assessment in mathematics brings both opportunities and challenges. Regardless, as technology becomes an integral part of mathematics assessment, a fundamental principle lingers: Assessment should focus on making the significant measurable rather than allowing what is measurable to define what is significant. This guideline, formulated over 30 years ago by the National Research Council Mathematical Sciences Education Board (1993), remains essential for any assessment practice. This is closely related to validity.

3 Validity of integrating digital tools in assessment through technology

Validity is a central concept in educational assessment research. Initially, it was defined as the extent to which a test measures what it is supposed to measure (Ruch, 1924). More contemporary frameworks on validity emphasise an argumentative validating process regarding the appropriateness of test score interpretations and uses (Kane, 2006). A definition aligned with this argumentative approach is outlined in the Standards for Educational and Psychological Testing (the Standards) (AERA et al., 2014). According to the Standards (AERA et al., 2014), *validity* is referred to as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (p. 11). Thus, validity involves gathering relevant evidence to provide a sound scientific basis for the proposed interpretations and uses of test scores.

In this chapter, we adopt a pragmatic view of validity, starting with the definition provided in the Standards (AERA et al., 2014). This pertains to the extent to which a particular type of mathematics item validly reflects relevant mathematical competence in an assessment situation, and whether it is reasonable to use such test scores to assess mathematical competence. Furthermore, it is unreasonable to conceptualise the validity of an assessment without considering it in relational terms. The validity of an assessment must, consequently, account for what is being assessed concerning the learning opportunities provided and the aims of the assessment activities. For example, to ensure a valid test situation, there must be constructive alignment (Biggs, 1996) between technology use and the preceding learning activities (Sangwin, 2013; Stacey & Wiliam, 2013). Therefore, the validation process (Kane, 2006) includes three significant components: what is assessed, what is intended to be assessed (the constructs), and how students are provided opportunities to learn what is assessed.

Regarding what is assessed, the focus of the assessment may centre on mathematical knowledge, digital skills, or sometimes both (Jankvist et al., 2021). However, it is essential to ensure that assessments remain focused on the relevant mathematics, minimising the influence of irrelevant technical skills (Drijvers, 2018; OECD, 2023a; Winnberg, 2025). Additionally, when students have access to digital technologies, it is especially important to define the intended constructs clearly (Jankvist et al., 2021; Stacey & Wiliam, 2013). The validation process thus evaluates the proposed uses and interpretations of test scores in relation to these constructs. Finally, it is equally important to investigate how students have been provided with opportunities to learn what is being tested, ensuring that the technical aspects of the items do not overshadow the focus on mathematics.

Examining the validity of test score interpretations and uses from a validity threat perspective (Messick, 1989) is also significant – particularly in the context of the technical and conceptual issues associated with digital assessments (Drijvers & Gravemeijer, 2005; Jankvist & Misfeldt, 2015; Stacey & Wiliam, 2013). For instance, if an assessment places excessive emphasis on students' technical skills that are only partially related to mathematical competence, the validity of its interpretations and uses may be compromised. Messick (1989), a seminal reference in validity theory, identifies two threats to valid inferences based on assessment outcomes: construct under-representation and construct-irrelevant variance. *Construct under-representation* occurs when an assessment fails to measure the intended constructs fully. This can happen, for example, when students' problem-solving strategies are assessed using multiple-choice questions, which are unlikely to capture their problem-solving processes. Construct-irrelevant variance, on the other hand, arises when an assessment measures aspects unrelated to the intended constructs. For example, if the digital tools provided are challenging for students, it becomes unclear whether low test scores reflect a lack of mathematical competence or difficulties with the digital tools. These validity-threat issues are central for anyone using assessments involving technology (Bennett, 2002). They are particularly significant in the context of PISA, a major global influence in education, as its results impact policy decisions in individual countries (Sollerman, 2019).

Our analysis will employ an instrumental approach perspective (Artigue, 2002; Guin & Trouche, 1998; Rabardel, 2002). Following Drijvers et al. (2010), this theoretical framework is well-suited to exploring the interconnected relationships between digital tool use and mathematical learning.

4 An instrumental approach

This study employs a theoretical approach perspective rooted in Vygotsky's (1978) theories on learning mediated by tools, namely, an instrumental approach. Within this perspective, some key distinctions exist between an artefact and an instrument. An *artefact* is a material or abstract object a subject uses to perform a specific activity with a particular intention. An *instrument*, however, is seen as the result of the subject's interaction with the artefact, combining the artefact with the user's cognitive utilisation schemes (Rabardel, 2002). Consequently, an artefact does not initially serve as an instrument; it becomes an instrument through the complex process of instrumental genesis, in which the subject develops relevant utilisation schemes (Verillon & Rabardel, 1995).

The process of instrumental genesis involves two interrelated progressions: one directed toward the artefact (instrumentalisation), and the other toward the subject (instrumentation) (Artigue, 2002; Rabardel, 2002; Trouche, 2005). *Instrumentalisation* pertains to the development of the artefactual components of the instrument and how the subject adapts the artefact's functions for specific uses (Rabardel, 2002). *Instrumentation*, on the other hand, involves the development of utilisation schemes and instrument-mediated actions, focusing on how the subject adapts to the artefact (Rabardel, 2002). The concept of instrumental genesis was later expanded by mathematics education scholars interested in technology use and learning (e.g. Artigue, 2002; Drijvers & Gravemeijer, 2005; Trouche, 2005).

Utilisation schemes are central to instruments. Following the reasoning of Drijvers and Gravemeijer (2005), utilisation schemes are categorised into usage schemes and instrumented action schemes. Usage schemes are basic schemes directly related to the artefact, such as the sorting function of a spreadsheet tool. An experienced user may apply a sorting scheme effortlessly, without considering the data relocations. In contrast, an inexperienced user must manage both the technical and conceptual aspects simultaneously and might feel anxious about the sudden data rearrangements in the spreadsheet tool. Instrumented action schemes develop from elementary usage schemes and are conceptualised as stable mental organisations where both technical and conceptual aspects interact during instrumental genesis. Whereas usage schemes directly relate to the artefact, instrumented action schemes connect the tool to carrying out transformations on the mathematical objects at stake (Drijvers & Gravemeijer, 2005). The key elements of an instrumented action scheme include technical and conceptual benchmarks that guide the transformation of an artefact into an instrument by the subject (instrumental genesis), facilitated by related usage schemes. We find the distinction between usage schemes and instrumented action schemes and the conceptualisation of the latter as schemes that include an interplay between technical and conceptual aspects relevant to the three central issues of validity: regarding what is being measured, the measurement objectives (constructs), and the alignment between the learning and assessment situation.

In the following sections, we initially frame our study within the context of PISA. Secondly, we identify the usage schemes and instrumented action schemes that students need to develop to solve mathematical problems successfully using spreadsheet tools. Thirdly, we will examine the technical and conceptual aspects of these instrumented action schemes that students have been given opportunities to develop during the training section. Finally, we will discuss our findings in relation to the validity of the inferences made from the assessment results.

5 Methods

This section introduces the PISA assessment and the obtained data, then presents our approach to identifying and analysing relevant schemes.

5.1 PISA 2022 assessment

Mathematics was the focus domain in the PISA 2022 assessment, which resulted in the development of new mathematics items and a revised theoretical framework (OECD, 2023a). This theoretical framework aims to guide the assessment of 15-year-old students' knowledge and skills across the core areas of reading, mathematics, and science. The framework sets the foundation for evaluating students' competencies in a way that reflects real-world problem-solving and critical thinking. Furthermore, it emphasises *literacy* in the three core areas, focusing on content knowledge and students' abilities to apply this knowledge in unfamiliar contexts. Specifically, *mathematical literacy* refers to an individual's capacity to reason mathematically and to formulate, use, and interpret mathematics to solve problems in real-world contexts. It encompasses concepts, procedures, facts, and tools to describe, explain, and predict phenomena, helping individuals understand the role of mathematics in the world and make informed decisions as engaged and reflective citizens (OECD, 2023a). The framework also advocates for updated methods for assessing students' use of technology in mathematical problem-solving activities.

One of the newly introduced areas in mathematics aims to place less emphasis on routine calculations to reflect the increased role of computer calculation tools (e.g. spreadsheet tools) used in professional contexts (OECD, 2023a). Therefore, new items employing spreadsheet tools were introduced in the PISA 2022 mathematics assessment. These items are generally motivated by technological advancements, the need to broaden assessment constructs, and the aim to reflect the evolving nature of mathematics in the modern world by including realistic datasets (OECD, 2023a). However, introducing items employing calculation tools comes with a key challenge: ensuring that these items continue to assess mathematical literacy while minimising inferences from irrelevant computer skills (OECD, 2023a). Students are, therefore, trained to use the spreadsheet tool beforehand to mitigate the possible impact of irrelevant technical skills. This training section is identical for all three units (in PISA, several items clustered within the same context are referred to as a *unit*) that utilise a spreadsheet tool. The assumption is that this training will ensure students reach a basic level of proficiency with the tool (OECD, 2023b).

5.2 The constructs in PISA 2022: spreadsheet items

Each mathematics item in the PISA assessment is categorised into a main content area (change and relationships, space and shape, quantity, and uncertainty and data) and a main problem-solving process (formulate, employ, and interpret/evaluate). The items analysed in this chapter fall mainly within the content area of uncertainty and data and the problem-solving process of interpret/evaluate, as categorised by PISA. Regarding spreadsheet tools, the focus on uncertainty and data involves understanding variation. This includes recognising how data is quantified and interpreting relevant information from tables. The process of interpret/evaluate is linked to computational models and encompasses the ability to analyse mathematical results and apply these insights within real-world contexts. Additionally, the learning objectives of the items employing spreadsheet tools emphasise students' ability to apply techniques and arithmetic operations in spreadsheet formulas and to interpret data output. These items are argued to assess (the constructs) students' capacity to analyse variation, draw conclusions from data, and effectively employ calculation tools in real-world contexts (OECD, 2023a).

5.3 PISA 2022 data: spreadsheet items

In this study, we focus on the Forested Area unit (CMA161) that employs spreadsheet tools. This covers four items and a total of eight questions. The Forested Area unit involves a complex use of the spreadsheet tool. For instance, in one of the items (presented in Figure 3.1), students need to use the calculation tool in several steps (first, calculate the difference for a time period, then calculate the difference for another time period, and lastly, calculate the difference between time periods) to solve the problem. This level of complexity, involving using previous calculations in a final calculation, is not required for the other items employing spreadsheet tools. In those cases, tool use is less complex, with results derivable in, at most, two steps. This is why we deliberately chose to focus on item CMA161Q03 (see Figure 3.1).

When spreadsheet items are used in PISA, they are preceded by an introduction and training section designed to familiarise students with the tool functions. This training covers sorting columns, performing arithmetic operations between selected columns, and calculating the mean of a selected column. Instructions are presented in text form, with prompts encouraging students to try out the functions. This approach assumes that the training will help students focus on the intended mathematics instead of on the uses of the spreadsheet tool (OECD, 2023a). Unlike ordinary classroom activities, such as summative assessments, large-scale assessments do not allow educators to provide additional related tool use information.

The left-hand side of Figure 3.1 first includes an expandable instruction about how to use the spreadsheet. This instruction was presented earlier in this unit's practice section, where students could train on using the spreadsheet tool. Second, there is an instruction, followed by a response component, where students select answers from drop-down lists. Students are prompted to consider two time periods, and the mathematical content involves changes in percentage points from one time period to the other. Columns E–G are used for the calculations performed. One way to productively use the calculation

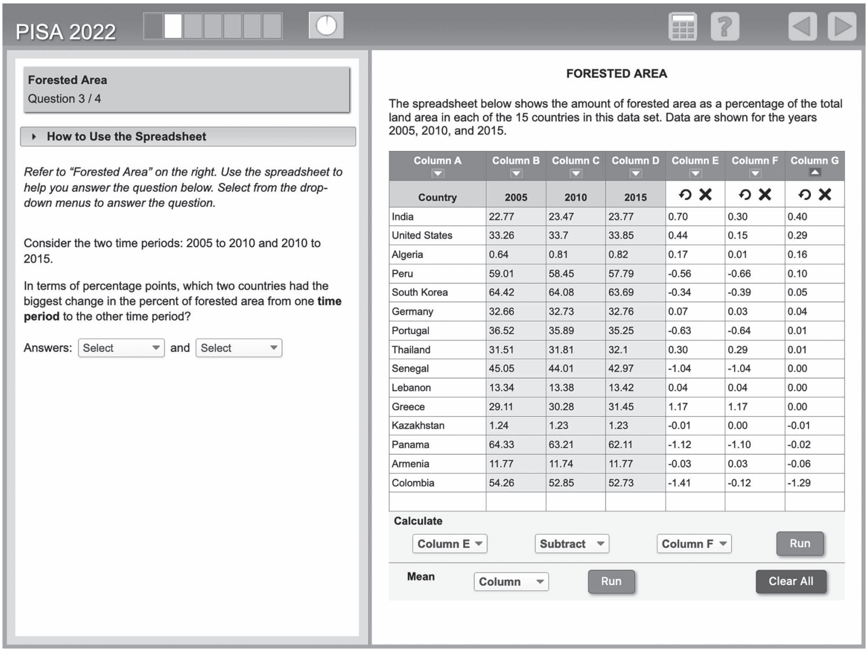


FIGURE 3.1 Example of a spreadsheet item: CMA161Q03. *Note:* 'Forested Area' is an item unit used in PISA 2022 and is one of the few items released for dissemination. For a full description of the introduction, practice section, instruction, and the four items in this unit (CMA161), see Annex C.

Source: Released items from the PISA 2022 computer-based mathematics assessment (OECD, 2023b).

tool is to (1) calculate Column C – Column B (results in Column E), (2) calculate Column D – Column C (results in Column F), and (3) calculate Column E – Column F (results in Column G) (see Figure 3.1). The right-hand side of the item displays the spreadsheet tool with the functionality of sorting spreadsheet data, calculating arithmetic operations between columns, and calculating the mean value for any selected column. However, the techniques employed in the spreadsheet tool differ in some ways from those of standard spreadsheet techniques. In the tool, columns are selected for calculations (arithmetic operations or computing the mean) within a separate component located at the bottom right (see Figure 3.1). In contrast, typical spreadsheets allow users to perform these calculations by selecting the column header or multiple cells within the column. Another distinction is that mean calculations in this tool are restricted to columns; the tool does not support row-wise mean calculations.

This study draws on student data from the PISA 2022 database (OECD, n.d.). Quantitative analysis was conducted using SPSS Version 29 to generate descriptive statistics for CMA161Q03, and a response matrix of student raw responses was investigated (see Section 6.3). Additionally, we identified usage schemes (see Section 6.1) based on the training section and hypothesised instrumented action schemes (see Section 6.2).

5.4 Approach to identifying and analysing schemes

Our approach consists of three stages. In stage 1, we identified the schemes at focus in the introduction for the spreadsheet item unit (Table 3.1). In stage 2, we formulated a hypothesised instrumented action scheme (Table 3.2) to solve the item. Stage 3, finally, consisted of relating the schemes that the students had an opportunity to learn to the required scheme to solve the item. In our analysis, we reviewed each key element and assessed whether the technical and conceptual aspects were covered in Table 3.1. Whereas stage 1 and stage 2 address research questions 1 and 2, respectively, stage 3 helped us relate both research questions to each other and to validity (research question 3), which is further elaborated in the discussion.

6 Findings

In this section, we present identified usage schemes for the training section, and the identified key elements in an instrumental action scheme for CMA161Q03, along with student data.

6.1 Identified usage schemes for the practice section

From the training section, we identify three usage schemes relevant to solving the items employing spreadsheet tools: the sorting scheme, the arithmetic calculation scheme, and the mean calculation scheme (Table 3.1). These schemes include technical actions (e.g. pushing a sorting button or selecting columns and arithmetic operations) for which the students receive training. They also encompass conceptual aspects (e.g. realising that the data is sorted in ascending order), with the information provided to explain (e.g. 'Note that all columns will sort based on the way any column is sorted').

6.2 Instrumented action scheme for evaluating change in percentage

Following the reasoning of Drijvers and Gravemeijer (2005), *instrumented action schemes* are seen as consistent and profound mental schemes constructed from fundamental usage schemes as a result of successful instrumental

TABLE 3.1 Identified usage schemes that students were given opportunities to develop before working with the spreadsheet item unit

Usage scheme	*Definition*	*Comments*
Sorting scheme	Data can be sorted with a sorting scheme. When the sorting button is pushed (technical), the user realises (conceptual) that the data in the spreadsheet is sorted in ascending order. The user also acknowledges (conceptual) that all the columns are sorted based on how any one column is sorted.	In the practice part of the unit, the student is prompted to use the spreadsheet tool to sort a column. The instruction is to click on the symbol in columns B, C, or D to sort that column in ascending order. A pop-up window appears if the student is inactive (>30 sec.) during the practice unit. If another period of inactivity (>30 sec.) passes, an animation shows how to perform the action.
Arithmetic calculation scheme	Data can be calculated with an arithmetic calculation scheme. When the calculation tool is used, the user selects (technical) any two columns from drop-down lists, an operation (adding, subtracting, multiplying, or dividing) from the middle drop-down list, and clicks (technical) on 'Run'. The user realises (conceptual) that the results from the selections are presented in the first available column to the right of the spreadsheet.	It is also possible to develop more complex arithmetical calculation schemes, such as using an operation applied to any two previous calculations (see Figure 3.1). However, this is not covered in the training section.
Mean calculation scheme	Calculating a column's (arithmetic) mean value can be performed with a mean calculation scheme. When the mean calculation tool is used, the user selects (technical) any column of the spreadsheet from a drop-down list and clicks (technical) on 'Run'. The user realises (conceptual) that the results from the operation are displayed in the cell below the selected column.	Not applicable to item CMA161Q03.

Note: The usage schemes, which are fundamental schemes directly related to the artefact, were identified from the techniques and related information that students are trained on in the practice section of the unit, for example, sorting data by pushing the sorting button. The schemes consist of technical and conceptual aspects, highlighted in the table.

TABLE 3.2 Hypothesised instrumented action scheme that students need to develop

Key elements	*Comments*
Step 1	
• Understanding the concept of change and percentage points	Conceptual, related to the mathematical problem at stake
• Knowing that the spreadsheet tool can be used for calculating differences between two given time periods	Conceptual, tool-related: trained arithmetic calculation scheme
• Knowing (from the practice, or applying the instruction) how to select columns and operations from the arithmetic calculation tool	Conceptual, tool-related: trained arithmetic calculation scheme
• Selecting the relevant columns and operation twice (e.g. column C – column B, and column D – column C), and clicking 'Run'	Technical, tool-related: trained arithmetic calculation scheme
Step 2	
• Understanding the concept of change between changes (the difference between two changes in percentage points)	Conceptual, related to the mathematical problem at stake
• Understanding that the relevant operation in the calculating tool can be applied for previous *results* (e.g. column F – column E), that is, *not only for given data*	Conceptual, tool-related: *not trained*; performing calculations on previous results obtained by the students; also conceptually related to the mathematical problem at stake
• Selecting relevant output columns and arithmetic operation (e.g. column E – column F) and clicking 'Run'	Technical, tool-related: trained arithmetic calculation scheme
• Knowing that sorting data will be useful	Conceptual, tool-related: trained sorting scheme
• Sorting the data output (column G)	Technical, tool-related: trained sorting scheme
• Interpreting the data in the given context (the biggest change)	Conceptual, related to the mathematical problem at stake

Note: The key elements are considered interconnected and co-evolving in the two steps.

genesis. We identify technical and conceptual key elements in a hypothesised instrumented action scheme for item CMA161Q03 in Table 3.2.

6.3 Student results for CMA161Q03

In the PISA assessment, the student proficiency scale for mathematical literacy is constructed based on item response theory (IRT) modelling approach. The scale for mathematics was first established in 2003, when mathematics was the focus domain, and set to an OECD average of 500 points, with a standard deviation of 100 points (OECD, 2024). For PISA 2022, the threshold score for achieving level 6 – the highest level of mathematical literacy – was 670 points. The third item in the Forested Area unit (Figure 3.1) has an item difficulty of 840 points (for full credit), making it one of the most challenging mathematics items in the PISA 2022 assessment.

The item CMA161Q03 is a partial credit, automatically scored item. Partial credit is awarded for responses where one of the selections is correct (while the other is incorrect or missing). Full credit is given when both correct countries are identified (key: India and Colombia, in any order). In the PISA 2022 dataset (OECD, n.d.), students' raw responses can also be obtained through the variables DMA161Q03RA (first selection) and DMA161Q03RB (second selection) associated with CMA161Q03. The two responses are selected from drop-down lists where the countries are displayed in alphabetical order.

A total of 20,474 students from OECD countries were assigned the item unit Forested Area in the adaptive test format of PISA 2022, where more challenging items are typically distributed based on students' prior results (6,418 OECD students were assigned the linear test format).

We intentionally focus on the students enrolled in the adaptive test format, as they are more likely to master the targeted mathematical content.

The item CMA161Q03 was quite difficult; 10% (2,056) of the students received full credit, 39% (7,973) partial credit, and 51% (10,445) no credit. OECD (2023b) argues that partial credit 'requires doing the same work that is needed for a full-credit response' (p. 391). In our interpretation, students who received partial credit had almost accomplished step 2 in the instrumented action scheme, but with an incorrect interpretation of the final data in column G. However, analysing the partial credit responses, 90% of students' (7,189) answers are likely derived from performing not more than step 1. Some of the most frequent answers in the partial credit category are Colombia (key) combined with Greece, or Senegal – a quite unreasonable selection if column G is appropriately generated (the difference is 0.00; see Figure 3.1). Moreover, only 10% (784) of the partial credit responses are likely to be derived from an incorrect interpretation of the data in step 2 (answers with the greatest absolute differences in column G: India [key] and United States, or Colombia [key] and Armenia; see Figure 3.1).

Our analysis of student responses connected to the hypothesised instrumented action scheme provides arguments in favour of the two steps – where the second step presents a conceptual challenge that a minority of the students manage to complete (see Table 3.3).

6.4 Instrumented action scheme related to the practice section

The usage schemes and the instrumented action schemes are related in the following way. While the usage schemes directly relate to the artefact (the spreadsheet tool), the more complex instrumented action schemes comprise

TABLE 3.3 Student responses associated with the two steps in the hypothesised instrumented action scheme

Response category	*Responses associated with students having completed step 1, not step 2*	*Responses associated with students having completed step 2*	*Students*	
			n	*%*
No credit			10,445	51.0
Partial credit	7,189	784	7,973	38.9
Full credit		2,056	2,056	10.0
Total	7,189	2,840	20,474	100

Note: The 784 students awarded partial credit likely completed step 2, except for correctly interpreting the data within the given context (last bullet point in step 2; Table 3.2).

foundational elements in related usage schemes. One can, therefore, consider the key elements in the instrumented action schemes as technical and conceptual milestones that guide the user's transformation of the spreadsheet tool into an instrument to solve the type of problem, as in CMA161Q03.

From Table 3.1, we conclude that the arithmetic calculation and sorting schemes are central for item CMA161Q03. While these usage schemes are essential for instrumental genesis, a key element in the corresponding instrumented action scheme for evaluating changes in percentage (see Table 3.2) – the second bullet point of step 2 – is not adequately addressed in the training section. Consequently, students must grasp that the calculation tool can also be used to process results they generate themselves. Importantly, they are not just working with the initial data provided, as was required for the previous items. Furthermore, the design of the tool makes this step significantly distinct in conceptualising its use. A clear visual difference exists between the column headings A–D (blue, representing the given data) and E–G (orange, representing students' calculations). When using the tool, it is insufficient to focus solely on columns A–D; students must recognise that working with columns E–G, which previously represented outcomes, is essential. Herein lies the complex process of instrumental genesis, where usage schemes are developed into more complex instrumented action schemes. We argue that this aspect of instrumentation presents a significant conceptually demanding facet related to tool use and the mathematical problem at stake.

Furthermore, if this conceptual element were straightforward, more students would likely appear in the partial credit category, representing those who nearly completed step 2 (see Table 3.3). This finding points to a possible misalignment between the training and what students are expected to perform with the spreadsheet tool. From PISA's (OECD, 2023a) point of view, the significant difficulty of items similar to CMA161Q03 should relate to the intended mathematics (first bullet point in step 2). However, since this complexity seems to relate to the second bullet point of step 2 for many students, a validity threat is likely present.

7 Discussion and conclusions

This chapter aimed to investigate what is assessed when digital technologies are integrated into large-scale assessments and how this influences the validity of inferences made from test scores. This aim was operationalised by investigating student data along with technical and conceptual aspects that students need to develop – and what learning opportunities they have been given in a training section before engaging with the test items.

The study findings suggest that students are primarily trained in techniques when introduced to spreadsheet tools. However, considering an instrumental approach, for the tools to become an instrument for the students, they need to develop relevant utilisation schemes in the complex process of instrumental genesis (Rabardel, 2002; Verillon & Rabardel, 1995). This process, which involves the intertwined and co-evolving development of technical skills and conceptual understanding (Drijvers & Gravemeijer, 2005), can be both time-consuming and challenging. Consequently, it raises questions about the feasibility of such development occurring within a one-moment test situation. Also, a dominant focus on digital skills might weaken the relationship between procedural and conceptual understanding (Jankvist et al., 2021). Although the investigated spreadsheet tool is more straightforward than typical spreadsheet applications, our findings highlight the need for more time and practice to develop instrumented action schemes and instrumental genesis. Such issues are likely to become even more pronounced with the introduction of more complex digital tools. This raises a crucial concern about the feasibility of expecting instrumental genesis to develop from a brief introduction and training section, which has direct implications for assessment validity.

The constructs of the items analysed in this chapter are presented in Section 5.2. Concerning validity (Kane, 2006), a reasonable chain of reasoning is as follows: The proposed score interpretation assumes that students have sufficient technical skills, developed through appropriate training, to productively use the spreadsheet tool in solving the mathematical problems in question. Additionally, the constructs appear to align with the assessment situation at first glance. As a result, the item scores can be interpreted as valid indicators of students' mathematical literacy, as they likely represent the intended constructs. Related to the instrumental approach, since the instrumented action scheme builds on the trained usage schemes and the additional conceptual elements align with the item's learning objectives, the item seems valid for assessing these learning goals.

However, scrutinising these validity arguments in light of the findings, we identify potential threats to the interpretations and uses of test scores.

While students were primarily trained in technical skills, the conceptual aspects of tool use (e.g. bullet point 2 in step 2, Table 3.2) may contribute to construct-irrelevant variance (Messick, 1989), undermining assessment validity. Additionally, the simplified spreadsheet tool differs significantly from typical spreadsheet uses, potentially introducing an instrumental distance (Haspekian, 2005). In this case, students are expected to handle a tool quite different from the contexts they are accustomed to – such as the standard spreadsheet applications used in classrooms – which creates challenges in integrating the tool into the assessment. To address these concerns, it is crucial to align the use of digital tools in large-scale assessments with prior learning activities (Sangwin, 2013; Stacey & Wiliam, 2013), ensuring that students have opportunities to develop both technical and conceptual understanding. These findings are significant for large-scale assessment practices, underscoring the need to integrate digital tools thoughtfully into learning situations before assessments while accounting for potential validity threats. Therefore, to ensure a productive use of digital tools in assessment situations, it is recommended that students have sufficient practice opportunities in problem-solving contexts prior to assessment. This recommendation applies even to simplified versions of a digital tool, as unforeseen technical and conceptual challenges can arise, making it essential to familiarise students with its functionalities in advance.

A similar recommendation applies to those using and interpreting results from large-scale assessments involving digital technologies. It is essential to critically reflect on how mathematics is assessed and how digital technologies are integrated into this process. Students' digital habits and access to technology can vary significantly across countries, leading to differing levels of familiarity with technology. Consequently, simplified versions of digital technology are often used in large-scale assessments. However, as shown in our study, these simplifications can introduce challenges for some students, which must be carefully considered when analysing, interpreting, and comparing results.

Finally, it is important to acknowledge the limitations of our study. The analysis was restricted to the item responses and the hypothesised utilisation schemes. While PISA data provides valuable insights, it does not capture the specific challenges students encounter when interacting with digital tools. To address this limitation, qualitative data from comparable assessment contexts could offer deeper insights and complement our findings. Notwithstanding this shortcoming, our study contributes to a deeper understanding of validity issues in large-scale assessments that incorporate digital technologies.

Reference list

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, *7*(3), 245–274. https://doi.org/10.1023/A:1022103903080

Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25–45). Springer Netherlands. https://doi.org/10.1007/978-94-007-4629-9_3

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, *1*(1), 1–24. https://doi.org/10.1002/9780470712993.ch11

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://doi.org/10.1007/BF00138871

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et Évaluation En Éducation*, *41*(1), 41–66. https://doi.org/10.7202/1055896ar

Drijvers, P., & Gravemeijer, K. (2005). Computer algebra as an instrument: Examples of algebraic schemes. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators: Turning a computational device into a mathematical instrument* (pp. 163–196). Springer US. https://doi.org/10.1007/0-387-23435-7_8

Drijvers, P., Kieran, C., Mariotti, M. A., Ainley, J., Andresen, M., Chan, Y. C., Dana-Picard, T., Gueudet, G., Kidron, I., Leung, A., & Meagher, M. (2010). Integrating technology into mathematics education: Theoretical perspectives. In C. Hoyles & J.-B. Lagrange (Eds.), *Mathematics education and technology-rethinking the terrain: The 17th ICMI study* (pp. 89–132). Springer US. https://doi.org/10.1007/978-1-4419-0146-0_7

Guin, D., & Trouche, L. (1998). The complex process of converting tools into mathematical instruments: The case of calculators. *International Journal of Computers for Mathematical Learning*, *3*(3), 195–227. https://doi.org/10.1023/A:1009892720043

Haspekian, M. (2005). An "instrumental approach" to study the integration of a computer tool into mathematics teaching: The case of spreadsheets. *International Journal of Computers for Mathematical Learning*, *10*(2), 109–141. https://doi.org/10.1007/s10758-005-0395-z

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Jankvist, U. T., Dreyøe, J., Geraniou, E., Weigand, H.-G., & Misfeldt, M. (2021). CAS from an assessment point of view. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalová, & H.-G. Weigand (Eds.), *Mathematics education in the digital age* (1st ed., pp. 99–120). Routledge. https://doi.org/10.4324/9781003137580-7

Jankvist, U. T., & Misfeldt, M. (2015). CAS-induced difficulties in learning mathematics? *For the Learning of Mathematics*, *35*(1), 15–20.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger Publishers. https://eduq.info/xmlui/handle/11515/34503

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

National Research Council (U.S.) Mathematical Sciences Education Board (Ed.). (1993). *Measuring what counts: A conceptual guide for mathematics assessment*. National Academies Press. http://ebookcentral.proquest.com/lib/sub/detail.action?docID=3376056

OECD. (2023a). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

OECD. (2023b). Released items from the PISA 2022 computer-based mathematics assessment. In *PISA 2022 results* (Vol. I). OECD Publishing. https://doi.org/10.1787/b7e7c683-en

OECD. (2024). *PISA 2022 technical report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

OECD. (n.d.). *PISA 2022 database*. OECD. www.oecd.org/en/data/datasets/pisa-2022-database.html

Pepin, B., Choppin, J., Ruthven, K., & Sinclair, N. (2017). Digital curriculum resources in mathematics education: Foundations for change. *ZDM*, *49*(5), 645–661. https://doi.org/10.1007/s11858-017-0879-z

Rabardel, P. (2002). *People and technology: A cognitive approach to contemporary instruments*. Université Paris. https://hal.archives-ouvertes.fr/file/index/docid/1020705/filename/people_and_technology.pdf

Ruch, G. M. (1924). *The improvement of the written examination*. Scott, Foresman & Co.

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

Sollerman, S. (2019). *Kan man räkna med PISA och TIMSS?: Relevansen hos internationella storskaliga mätningar i matematik i en nationell kontext* [Can we count on PISA and TIMSS? The relevance of international large-scale assessments in mathematics in a national context] [Dissertation, Stockholm University]. http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-167442

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. (Ken) Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 721–751). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Trouche, L. (2005). An instrumental approach to mathematics learning in symbolic calculator environments. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators* (Vol. 36, pp. 137–162). Springer-Verlag. https://doi.org/10.1007/0-387-23435-7_7

Verillon, P., & Rabardel, P. (1995). Cognition and artifacts: A contribution to the study of though in relation to instrumented activity. *European Journal of Psychology of Education*, *10*(1), 77–101.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Weigand, H.-G., Trgalova, J., & Tabach, M. (2024). Mathematics teaching, learning, and assessment in the digital age. *ZDM – Mathematics Education*, *56*(4), 525–541. https://doi.org/10.1007/s11858-024-01612-9

Winnberg, M. (2025). The use of spreadsheet tools in assessment: An instrumented technique perspective. *Nordisk matematikkdidaktikk, NOMAD [Nordic Studies in Mathematics Education]*, *30*(1), 83–102. https://doi.org/10.7146/nomad.v30i1.152932

Yerushalmy, M., Nagari-Haddif, G., & Olsher, S. (2017). Design of tasks for online assessment that supports understanding of students' conceptions. *ZDM*, *49*(5), 701–716. https://doi.org/10.1007/s11858-017-0871-7

4

HIGH-STAKES MATHEMATICS ASSESSMENT IN FINLAND

Thomas Vikberg

1 Introduction

At nine o'clock in the morning on Tuesday, 26 March 2019, 26,473 Matriculation Examination candidates sit in examination halls in general upper secondary schools all around Finland, ready to start their examination day. They boot up their computers from USB sticks into a digital exam system and connect to their school's server through a local network. In front of them is the first digital matriculation examination test in mathematics, with which they are expected to spend the next six hours answering. Technical problems occur for some candidates as their computers experience glitches, or simultaneous memory-intense use of calculator software freezes the computer, but by the end of the day, everyone has completed their test, and every answer is securely stored in an online assessment service.

The Matriculation Examination, as defined by the Act on the Matriculation Examination (2019a), assesses whether candidates have achieved the knowledge, skills, and maturity specified in the curriculum for general upper secondary education. Administered biannually in all general upper secondary schools, it provides general eligibility for higher education. As Finland's only national high-stakes examination, it is taken by approximately half of each age group after completing nine years of compulsory basic education and three years of secondary education. While the other half typically pursue vocational education and are eligible to participate, their involvement is rare due to the examination's alignment with general upper secondary education objectives. Additionally, adult learners and retakers contribute to the candidate population.

DOI: 10.4324/9781003533764-4

Since the spring of 2019, the Matriculation Examination has been fully digital, with mathematics being the final subject to make the transition. This chapter employs archival policy analysis reflecting Hill and Hupe's (2022) distinction between policy formation and policy implementation, with the following section focusing on the development of policy goals and decisions during the formation phase, while the following section examines how these policies were translated into practice during implementation. As noted by Hill and Hupe (2022), the distinction between these phases can be debated and is not clear-cut. However, in this chapter, a top-down perspective is adopted, emphasising that 'implementation takes place in institutional settings embedded with mandates of legitimate political authority' (Hill & Hupe, 2022, p. 54).

The purpose of this analysis is to understand the historical precedents, explore the phases of development, and provide context for the reform. Following these sections, the chapter includes an analysis of the first digital mathematics tests and presents key statistics on how the shift to digital tests has influenced students' choices of subjects in the examination.

The chapter draws on a range of primary sources, including government documents, policy papers, meeting minutes, reports, and other archival materials. Sources have been gathered by examining publicly available online resources, by requesting information from authorities, and by information gathering in the Matriculation Examinations Board's digital and manual archives under permit no. OPH-4046–2024. With many of these sources being in Finnish or Swedish, the official languages in Finland, they may not be accessible to an international audience, adding value to this exploration. By examining these materials, the chapter sheds light on the interplay between key actors and the critical decisions during the years leading up to the reforms, as well as the collaborative efforts that culminated in the first digital tests. It also illustrates one possible approach to organising and delivering high-stakes mathematical assessment digitally.

2 Rationale and foundations for the digital transition

2.1 Legislative context and stakeholder advocacy

By constitution, Finland is a republic with a unicameral parliament holding legislative power. Within the administrative branch of the Ministry of Education and Culture (before 2010, Ministry of Education) are the national authorities the Finnish National Agency for Education and the Matriculation Examination Board.

The Matriculation Examination Board, where the author has held an administrative position since autumn 2016, is an independent expert body appointed by the Ministry of Education and Culture. It is responsible for managing, organising, and implementing the Matriculation Examination.

The Board prepares test items based on the compulsory and nationally elective study modules for each subject, as specified in legislation. The Finnish National Agency for Education oversees the development of education from early childhood to adulthood, including issuing the national core curricula for general upper secondary education, guided by relevant laws and decrees.

The goal of developing strong information society skills in students was first established in the Government Decree on the General National Objectives of General Upper Secondary Education and the Distribution of Lesson Hours (2002). Mastering these skills was included in the subsequent National Core Curriculum (Finnish National Agency for Education, 2003). However, the continued reliance on pen-and-paper methods in the Matriculation Examination limited the extent to which these goals could be realised in classroom practices, as assessment practices influenced teaching and learning priorities.

In particular, one group was unhappy with the pace of change: the students. In its Target Program for 2008, the Union of Upper Secondary School Students (2007) demanded that Matriculation Examination candidates be given the opportunity to answer tests using a computer. According to the Union, this reform would allow candidates more time to focus on more essential aspects in the test situation. The Union further stressed that, within this reform, sufficient attention be paid to ensure the legal safety of candidates and the operating of the technology during the tests and in any potential exceptional situations, such as during power outages. The demand was reiterated in the Target Programmes of 2009 and 2010, with the addition that the first pilots to be tested could be in the written essay assignment in mother tongue tests (Union of Upper Secondary School Students, 2008, 2009).

Acknowledging the need for reform, the Ministry of Education (2008), in its regular Development Plan for Education and Research, noted that available technological capabilities were not utilised in the Matriculation Examination. The Development Plan recommended reviewing the development needs of the Matriculation Examination and proposed an investigation into the potential use of information technology tools in the examination process, alongside an analysis of the associated costs.

2.2 Building consensus around reform proposals

In 2009, a committee was appointed by the Ministry of Education to propose measures for the development of general upper secondary education. One of the tasks of the committee was to make proposals on how the Matriculation Examination should be developed. In its final report, the Committee proposed, among others, that information technology be gradually adopted in Matriculation Examination tests from 2014 onwards (Ministry of Education and Culture & Pirhonen, 2010). According to the report, for technology to be utilised in the examination, general upper secondary schools would need adequate and consistent technological capabilities, and it would be necessary to utilise technology in teaching and student assessment across all schools. According to the report, using digital technology in the Matriculation Examination would necessitate the development of the tests and determining its applicability in various subjects. This would include evaluating and improving test items and instructions while also considering candidate identification, security requirements, technological issues, candidate equality, and exam fraud risks.

The committee predicted that the cost for the proposed reform would be 10 million euros for development, followed by 1 million euros annually for support and maintenance.

At a subsequent round of public commenting, the proposal to incorporate information technology in the Matriculation Examination got mostly positive feedback by stakeholders. The proposal was strongly backed by the Union of Upper Secondary School Students (2011) as the issue had already become a promotion target for them in the upcoming government program negotiations (Union of Upper Secondary School Students, 2010). The Finnish Association of Teachers of Mathematics, Physics, Chemistry and Informatics (2011) was positive to the reform, as well as the Trade Union of Education in Finland (2011), which supported the proposal, provided that teachers get sufficient in-service training and tools and that the reform be made in sufficiently small steps, starting with curriculum reforms.

The biggest reservations were directed towards the timetable of the reform. In its statement, the Matriculation Examination Board (2011a) supported the proposed reform but stated that students should know at the beginning of their studies how their high-stakes final exams are conducted. This would mean that in order to meet the target of 2014, students starting their three-year studies in autumn of 2011 should know by then to some extent the details of how computers would be used in the tests. This would mean that this information should be ready in half a year, and therefore, planning of the reform should begin immediately.

2.3 Implementation framework and timelines

In April 2011, Finland held parliament elections. In the electoral system, no single party gains a majority, so to form a majority coalition government, parties elected to parliament undergo government negotiations to form a common program. The program of 2011 was long and detailed, with one of the agreed-upon goals to prepare for the gradual implementation of information and communication technology in the Matriculation Examination (Prime Minister's Office, 2011).

The following year, state funds were allocated for the Matriculation Examination Board to start to implement the reform. The Matriculation Examination

Board appointed a working group to prepare and propose necessary decisions. Following the advice from the working group (Matriculation Examination Board & Lokki, 2013), the Matriculation Examination Board (2013a) decided on the timetable for the reform. Unlike the original proposal of introducing digital tests in 2014, the Board decided to progressively increase the number of digital tests from autumn 2016 subject by subject. As the expectation was that mathematics would be the most difficult to organise digitally, as transforming the answering process from pen-and-paper to digital seemed challenging, these tests would be the last to go digital in spring 2019.

In 2015, the Finnish National Agency for Education (2016) introduced a new core curriculum, which significantly redefined how digital competencies were integrated into general upper secondary education. In mathematics, ICT skills were no longer treated as distinct from mathematical skills but were instead embedded within the framework of mathematical competence. The revised curriculum advocated the use of a range of digital tools, including dynamic mathematics software, symbolic computation software, statistical software, spreadsheets, word processing tools, and digital information sources, wherever appropriate. Moreover, it emphasised the importance of equipping students with the ability to critically assess the usefulness and limitations of these tools. To this end, each compulsory and nationally advanced mathematics study module was designed to include at least one learning objective explicitly focused on ICT skills.

These changes underscored the necessity of developing a testing medium capable of evaluating such digitally oriented competencies. For instance, in the study module Mathematical Analysis (MAB7), an objective became that students 'use technical tools to examine the continuity of a function and to define the extreme values of a bounded interval in application tasks' (Finnish National Agency for Education, 2016, p. 148). Similarly, the module Probability and Statistics (MAA10) set objectives for students to 'use technical tools to acquire, process, and examine digital data as well as to determine distribution parameters and calculating probabilities with the help of the distribution parameters provided' (Finnish National Agency for Education, 2016, p. 144). As the new curriculum came into effect for students beginning their studies in autumn 2016, those participating in the first digital tests in 2019 were already being taught with learning objectives that aligned more closely with the digital format.

3 Development and challenges in mathematics testing

3.1 Evolution of calculator policies

The first steps in the use of digital technology in mathematics tests were taken when the Matriculation Examination Board (1981) allowed programmable calculators in the Matriculation Examination. To prevent cheating, the calculators needed to be such that their memory could be wiped out before the test.

Over the years, regulations were amended so that calculators that could provide results in graphical form in addition to numerical form were allowed. Calculating the value of a function or expression, calculating basic statistical parameters, numerically solving an equation or system of equations, as well as numerical differentiation and integration, were acceptable operations. Calculators were not allowed to include algebraic or analytical operations handling symbolic expressions, such as symbolic differentiation and integration.

According to a new directive issued by the Matriculation Examination Board (2011b), starting from the examination of spring 2012, calculator regulations were amended so that all handheld function, graphical, and symbolic calculators were allowed in mathematics tests. This reform had consequences in how test items were formulated, as candidates could now be expected to possess a powerful digital device during tests.

It was soon realised that the reform made it difficult to assess some core mathematical competencies. The regulations were further amended from 2016 onward so that the tests were split into two sections, A and B, where no calculators were allowed in Section A (Matriculation Examination Board, 2013b). The logistics of the reformed tests was such that candidates started without a calculator, and after handing in the answer papers of Section A, they were handed out their calculator to be used for the rest of the test.

The gradual expansion of calculator use and the need to balance computational tools with core mathematical reasoning laid the foundation for the development of digital mathematics tests. As discussed earlier, the design of mathematics tests prior to the digital transition had already accounted for the possibility that candidates might have access to powerful digital tools. Therefore, the shift to digital tests did not introduce an entirely new situation but rather built upon existing considerations.

3.2 Designing the digital exam system

Work on the digital Matriculation Examination started in 2013. After an initial year of exploration and research on digital exam systems for high-stakes testing, the Matriculation Examination Board decided to develop its own environment (Lattu, 2014). The following requirements were set for the exam system:

1. Candidates work during tests using a computer that offers the same diverse set of software to everyone.
2. Tests are taken on the candidate's own computer, without the need for any permanent modifications (e.g. software installations).

3. Requirements for the devices used in the test are so low that they are easy to acquire and are available throughout Finland.
4. During the test, candidates' access to the computer's functions can be restricted.
5. Candidates' computer usage can be technically monitored during the test.
6. The exam system can handle exceptional situations (e.g. computer or network failure, power outage) with the help of the test supervisors without requiring them to have special technical expertise.
7. Candidates' test answers cannot be lost under any circumstances.

From the outset, it was evident that meeting these requirements would preclude the exam setup from being a straightforward plug-and-play solution. The chosen system design was to provide candidates with a fully functional Linux operating system. A Live Linux system is a Linux distribution that can be booted and executed directly from a removable storage device, such as a USB drive. These systems operate entirely in a computer's memory (RAM) and leave the host system unchanged, unless specifically configured to make modifications. This setup allowed the candidates' insecure hard drives to be bypassed, and the operating system could include various pre-installed software and files. Consequently, each candidate was ensured a uniform environment, regardless of the computer they brought to the examination, with no installations required on the candidates' computers.

The quality of internet connection in schools was highly variable at this stage, and to mitigate risks, it was decided that no internet connection would be used during tests; all network traffic would take place in a local network within the school premises.

By early 2015, an early prototype of the exam system, named *Abitti* – derived from the German *abitur* (upper secondary students) and the Finnish *bitti* (basic unit of information in computing) – was launched (Lattu, 2015). The introduction of Abitti marked a significant milestone, enabling teachers and students to engage directly with the digital testing environment. This hands-on access shifted teachers from passive recipients of project updates to active participants, fostering a deeper understanding of the system and its potential for high-stakes assessments.

In autumn 2016, the first digital Matriculation Examination tests were successfully conducted in German language, geography, and philosophy.

3.3 Navigating concerns

In a letter to the Matriculation Examination Board in the autumn of 2016, the Finnish Association of Teachers of Mathematics, Physics, Chemistry and Informatics (2016) voiced concerns about their ability to adequately prepare students for the digital Matriculation Examination. In its response, the Mathematics Division of the Matriculation Examination Board (2016) explained their view on the upcoming digital mathematics examination, emphasising the importance of preserving teachers' pedagogical freedom to utilise the tools they found most effective in their teaching practices. A recurring question at that time was why a single calculator software was not chosen for the exam system. The Division argued that mandating one specific calculator would position the Board as an authority dictating the choice of tools in education. Furthermore, adopting a single calculator software could lead to a negative washback effect, compelling educators to conform to the mandated tool regardless of its suitability for their pedagogy. Additionally, ensuring standardised access to calculator software for all candidates addressed equity concerns by resolving a prior issue where not all students could afford advanced handheld calculators.

To address further concerns raised by teachers, the Matriculation Examination Board (2017a) collected feedback through multiple channels. An online survey conducted in January 2017 sought input from mathematics, physics, and chemistry teachers on issues they found concerning in the shift to digital assessments. Additionally, three workshops held during the same period gathered further questions and feedback.

While specific response numbers for the survey are unavailable, the initiative ensured representation by targeting educators from multiple disciplines. Teachers raised several subject-specific concerns, particularly related to mathematics education (Matriculation Examination Board, 2017a). For example, they highlighted challenges with the use of ICT tools, including the compatibility of calculator software with test answer production and the presentation of mathematical notation. Many expressed apprehension that students trained to use certain tools might face difficulties adapting to the requirements of digital tests.

The collected feedback was synthesised into a report addressing 31 frequently asked questions, which was published by the Mathematics, Physics, and Chemistry Divisions of the Matriculation Examination Board (2017a). This report aimed to clarify the rationale behind various decisions, such as the selection of permissible tools, and to reassure educators about the alignment of digital assessment practices with established pedagogical objectives.

As the first digital mathematics tests approached, concerns among students and teachers intensified in autumn 2017. An article in *Helsingin Sanomat*, Finland's largest newspaper, highlighted frustration with the digital tools designated for the reform. Students criticised the software as clumsy and slow, questioning its imposition over traditional methods, which they found more effective (Valtavaara, 2017). These growing concerns prompted a mathematics teacher to lodge a formal complaint with the Parliamentary Ombudsman, citing inadequate preparation for the transition. Concurrently, a student petition calling for the postponement of the digital mathematics tests amassed

4,186 signatures (Pekkala, 2017; Lattu & Vikberg, 2018), while another petition was submitted to school principals and heads of education authorities advocating for a similar delay (Tolvanen, 2017).

Despite the concerns raised, the reform moved forward as planned. By this stage, altering the timeline was deemed impractical and unfair to students and teachers who had already adapted to the new curriculum and incorporated information and communication technology into their teaching and learning practices (Lattu & Vikberg, 2018). Heads of education authorities from five major cities emphasised the importance of adhering to the schedule, cautioning that delays would disrupt the comprehensive development of general upper secondary education (Heikkonen, 2017). The Parliamentary Ombudsman (2017) determined that the preparation and communication surrounding the digital mathematics tests ensured equitable treatment of candidates and furthermore concluded that the digital medium did not disadvantage students in higher education admissions, supporting the decision to maintain the reform timeline.

Still, concerns continued amongst teachers, parents, and students who saw that digital tests negatively changed fundamental aspects of learning mathematics. Specifically, apprehensions were raised about the impact on the process of writing detailed, open-ended answers with intermediate steps, as well as the time required to master digital tools potentially detracting from learning core mathematical concepts (see, for example, opinion pieces in *Helsingin Sanomat*: Seppänen, 2018; Väisälä, 2018; Mäntylä, 2018).

3.4 Writing mathematical notation with a computer

A prominent concern among teachers and students revolved around the method of writing mathematical text in future matriculation tests, specifically in Section A, where scientific calculators with their own formula editors would not be permitted (The Finnish Association of Teachers of Mathematics, Physics, Chemistry and Informatics, 2016).

The initial approach was to identify an existing formula editor suitable for use in the Matriculation Examination, thereby avoiding the need to create a specialised, exam-specific tool. However, the lack of a suitable editor and the constraints of the development timeline led to the decision to design a new editor (Vikberg, 2017). During the early phases of planning, trials were conducted in several general upper secondary schools, involving students with varying levels of computer proficiency. These trials revealed that students could achieve an adequate typing speed for mathematical notation after just 30 minutes of training, indicating a rapid learning curve. The ease with which students adapted to the trial tools suggested that mastering the editor yet to be developed would also facilitate the use of other formula editors in the future, alleviating concerns about creating a proprietary tool (Vikberg, 2017).

On 7 May 2017, the Matriculation Examination Board (2017b) released a version of the Abitti exam system featuring an editor capable of incorporating mathematical notation and allowing screen captures within the answer field. Additionally, the system introduced functionality to restrict the use of advanced scientific calculators during tests. While ongoing improvements to the exam system were planned, this release marked a significant milestone, as all essential features required for creating and completing practice tests in mathematics were now operational.

Further, the Matriculation Examination Board tasked a pair of teachers with producing digital answers to a prior mathematics test item using various tools and approaches. The test item was a 12-point question, and representatives of the Mathematics Division assessed these answers, providing score ranges and detailed comments on their strengths and weaknesses (Vikberg, 2018). One solution, for instance, effectively utilised calculator functions but lacked explanation, prompting the comment 'The solution lacks an explanation of what the calculations aim to achieve. The reader must guess how the computations relate to the problem. Why do these calculations lead to the answer?' Such shortcomings led to a score range of 6–8 out of 12 points. Figures 4.1 and 4.2 represent two 12-point solutions by Hellsten and Rahikka (2018), highlighting the many possibilities of successfully demonstrating mathematical competence in the tests.

The academic year of 2018–2019 saw substantial use of the newly implemented practice tools. A particularly notable tool was the formula editor demo page, which received around 40,000 hits on regular weekdays. This increasing usage was closely monitored with mixed feelings by the Matriculation Examination Board, as the system's growing popularity highlighted its critical role in the success of the digital Matriculation Exam (Lattu & Vikberg, 2018). Despite its widespread use, the Board stressed that the Abitti exam system was designed as a practice environment for the examination. Teachers were encouraged to explore a variety of digital tools in their pedagogy, as, for example, the formula editor demo page was primarily intended as a demonstration tool and was far from a comprehensive educational service.

4 Digital matriculation examination in mathematics

4.1 The first digital mathematics tests

In the spring of 2019, all Matriculation Examination tests in Finland were conducted digitally. The examination consisted of 42 tests in different subjects and syllabus, with mathematics offered in two separate tests: Advanced Level and Basic Level. The mathematics tests on 26 March saw a typical amount of minor technical issues, with a total of over 30,000 computers being used over a test time of 6 hours (Matriculation Examination Board, 2019a).

Olkoon funktiot $f(x):=2\cdot e^{-x}$ ▸ *Valmis* ja $g(x):=x^2\cdot e^{-x}$ ▸ *Valmis*.

Kuvasta nähdään, että leikkauskohtien välissä funktion **f** kuvaaja on funktion **g** kuvaajan yläpuolella. Voidaan siis määrittää janan pituudelle funktio

$h(x):=f(x)-g(x)$ ▸ *Valmis*

Funktio **h** saavuttaa suurimman arvonsa kohdassa $\mathrm{fMax}(h(x),x)$ ▸ $x=-(\sqrt{3}-1)$

Kohta on leikkauspisteiden välissä, sillä $-(\sqrt{3}-1)$ ▸ -0.73, joka selvästi on leikkauspisteiden välissä.

Lasketaan funktion suurimman arvon tarkka arvo ja kaksidesimaalinen likiarvo

$h(x)|x=-(\sqrt{3}-1)$ ▸ $(2\cdot\sqrt{3}-2)\cdot e^{\sqrt{3}-1}$

$h(x)|x=-(\sqrt{3}-1)$ ▸ 3.04

Vastaus: Janan pituuden maksimiarvo on $(2\cdot\sqrt{3}-2)\cdot e^{\sqrt{3}-1}$. Sen kaksidesimaalinen likiarvo on 3,04.

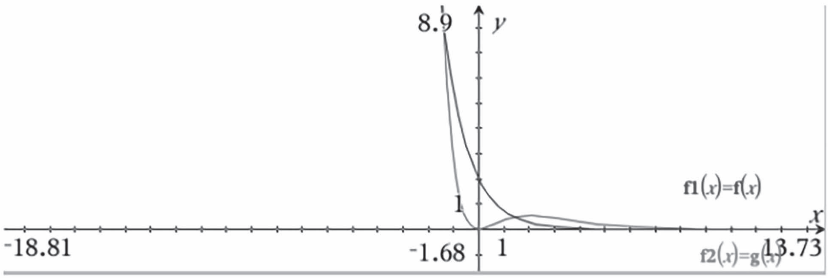


FIGURE 4.1 Digital answers to a previous mathematics test using screenshot from Texas Instruments TI-Nspire CAS.

All mathematics tests, including those conducted in spring 2019, are published after the test day on the Finnish Broadcasting Company's website (2024). Both the Advanced Level and Basic Level tests consisted of 13 items, of which the candidate had to answer 10. The tests were divided into three sections. Section A contained four mandatory items. Section B1 contained five items, of which the candidate had to answer three. Section B2 contained four items, of which the candidate had to answer three. All items were graded on a scale of 0–12, which meant that the maximum score for the test was 120.

To ease the transition to digital tests, the Matriculation Examination Board decided to maintain the structure of the mathematics tests unchanged during the reform process, with more significant alterations introduced in later phases (Matriculation Examination Board, 2022). This approach is evident in

Määritellään uusi funktio h(x)=f(x)-g(x).
Derivoidaan funktio (rivi 2).
Etsitään derivaattafunktion h' nollakohdat (rad 3).
Tutkitaan funktion h arvoa suljetulla välillä

$\left[-\sqrt{2}, \sqrt{2}\right]$ (rivi 1).

Funktio saa suurimman ja pienimmän arvonsa suljetun välin päätepisteissä
tai derivaatan nollakohdissa.
Huomataan, että suurin arvo saadaan kohdassa

$x = -\sqrt{3} + 1$, jolloin funktioiden välinen etäisyys on

$e^{\sqrt{3}-1}\left(2\sqrt{3} - 2\right)$ (rivi 8).

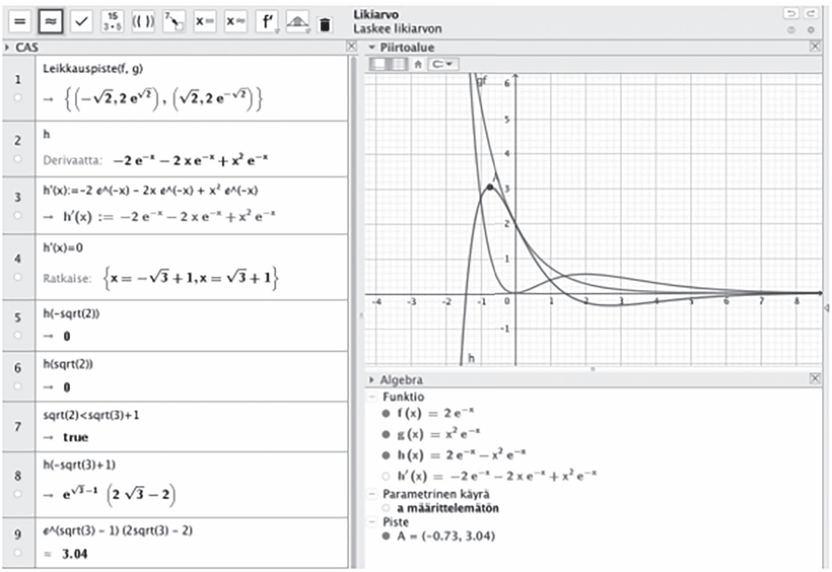


FIGURE 4.2 Digital answers to a previous mathematics test using text, formula editor, and screenshot from GeoGebra.

the test items of the spring 2019 in mathematics, where the shift to a digital medium had only a modest impact on the design of the test items.

In both Advanced Level and Basic Level tests, the first test item was a set of multi-choice questions, and the rest were open-ended items, meaning that candidates were expected to answer nine open-ended items. Both tests included items involving supplementary materials. These materials featured function graphs, and one item in both tests was based on the use of spreadsheet data. The Advanced Level test included an item with a GeoGebra file that allowed candidates to explore the problem in the question, but solving the task did not require processing the provided materials.

In Section A, candidates could utilise a basic calculator, KCalc, provided by the exam system. A candidate could answer the items in Sections B before submitting Section A if they so wished. After submitting, Section A's answers to that section could no longer be modified and all blocked programs (LibreOffice Calc, wxMaxima, TI-Nspire CAS CX Student Software, Casio ClassPad Manager, Logger Pro, Geogebra 5 and 6, 4f Vihko) became available. Several of these software could also be used to manipulate spreadsheet data. Additionally, the exam system incorporated multiple software tools for creating images and diagrams, including both bitmap and vector drawing applications. Additionally, over a transition period until the end of 2020, own handheld calculators could be collected and used after submitting Section A.

After the test, the Mathematics Division of the Matriculation Examination Board (2019b) published detailed evaluation criteria and examples to show how points were awarded based on individual merits. The general criteria emphasised that answers should demonstrate clear reasoning, include all necessary calculations or justifications, and present a correct final result. The assessment criteria considered the entire solution, including the initial steps, intermediate work, and final conclusion. Minor calculation errors that did not affect the overall task had a limited impact on the score, while significant errors in calculation or modelling could lead to substantial score reductions.

Since the first digital mathematics tests, the assessment criteria have evolved to include more detailed guidelines, particularly for awarding points for insightful visual representations, often created using calculator software (Matriculation Examination Board, 2024). However, the general criteria for calculator use remain unchanged: If a symbolic calculator is used to solve an item, this must be explicitly indicated in the answer. For items requiring interpretation and analysis, a calculator-generated answer alone is insufficient without additional justification. Nevertheless, results from a calculator are generally accepted for routine items and the routine components of larger answers. Examples of such items include simplifying expressions, solving equations, and calculating derivatives or integrals of functions.

4.2 Participation trends and insights

According to public financial statements, by the end of 2024, a total of 191,900 test submissions had been made in the digital mathematics tests by 165,808 individuals (Finnish National Agency for Education, 2020, 2025). Figure 4.3 presents the number of test submissions for the Advanced Level, while Figure 4.4 illustrates the test submissions for the Basic Level over a period of 10 years.

As shown by the data, the transition to digital tests led to an increased number of candidates of Basic Level who organised their studies so that they could participate in the last paper-and-pen test in 2018, as 15,846 candidates participated in 2017, 17,298 participated in 2018, and 12,552 participated

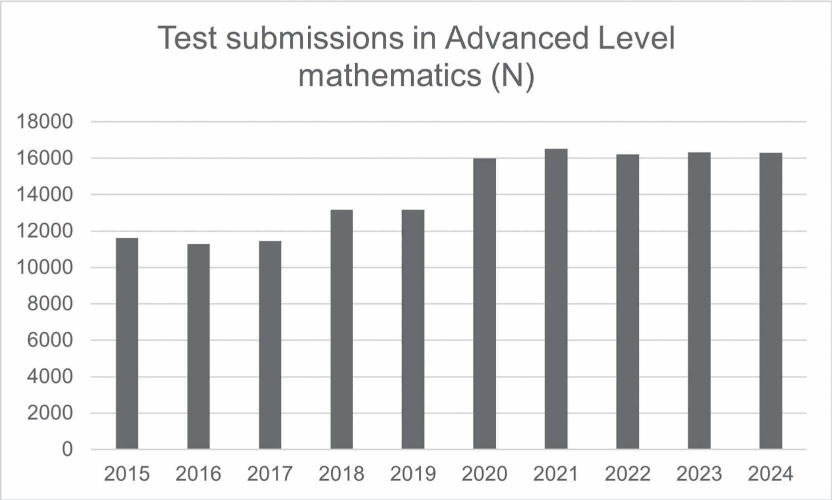


FIGURE 4.3 Test submissions in Advanced Level mathematics (*N*).

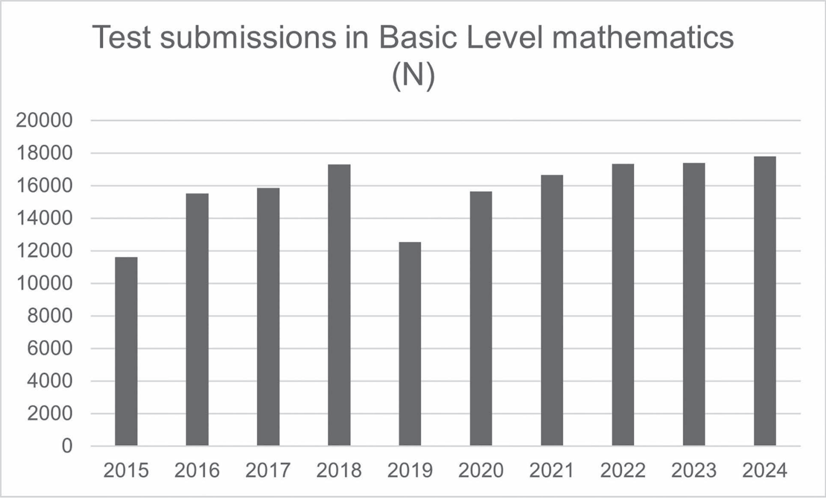


FIGURE 4.4 Test submissions in Basic Level mathematics (*N*).

Source: Finnish National Agency for Education (2020, 2025).

in 2019. However, such organisation of studies was practically very difficult for candidates taking the Advanced Level tests due to the extensive number of study modules required for participation in the test, so no particular change in candidates' behaviour can be seen.

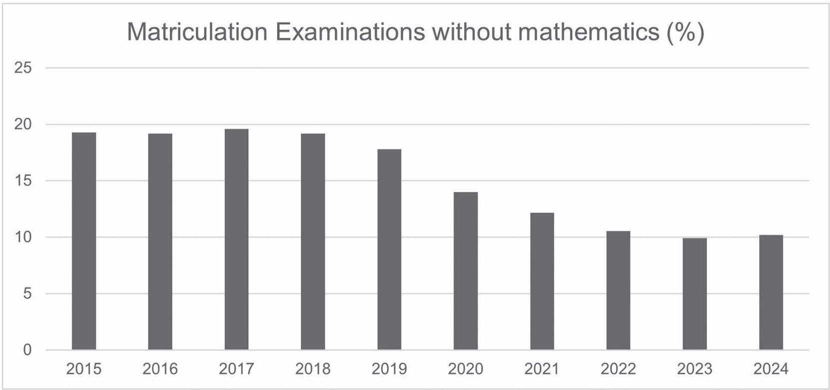


FIGURE 4.5 Candidates opting out of mathematics in the Matriculation Examination (%).

Mathematics is an elective subject in the Matriculation Examination, as only mother tongue and literature is a mandatory subject. Open data from the Matriculation Examinations Board (2025) illustrate in Figure 4.5 a significant decrease in examinations without mathematics as a completed subject over a ten-year period, with 19.3% in 2014 and 10.2% in 2023.

However, many factors influence participation in the tests, such as changes in the higher education admission system in Finland, which has increased the weight on the performance in mathematics from 2020 onwards (The Council of Rectors of Finnish Universities, 2018), and new legislation increasing the required number of subjects in the Matriculation Examination from four to five from 2022 onwards (Act on the Matriculation Examination, 2019). Consequently, no definitive conclusions can be drawn from the data, except that participation rates have not decreased since the digital tests began.

5 Discussion

The reform to digitise the Matriculation Examination in Finland was rooted in documented evidence of policy formation, which can be distinguished from the implementation phase of the policymaking process (Hill & Hupe, 2022). Archival records, including policy programs, committee reports, and official communications, highlight how a shared consensus emerged among key stakeholders that the traditional pen-and-paper assessment model was impeding the development of general upper secondary education. However, as Scott (1990) reminds us, official documents should be approached keeping in mind that they reflect the administrative routines and organisational priorities of their creators, serving as tools within broader policy and administrative processes highlighting the perspective for a critical examination of these records, recognising their role as shaped artefacts of policy formation.

The records further demonstrate that the reform was presented not merely as a technical adjustment but as an educational policy initiative designed to promote the integration of digital technology into general upper secondary education. The primary motivations for these measures appear to be the necessity for education to align with the digital progression of society and the need for more versatile test items and accompanying materials consistent with the upper secondary curriculum. By aligning assessment methods with the digital competencies required in contemporary learning environments, the reform aimed to drive a broader transformation in educational practices. Consequently, the digitisation initiative can be interpreted as a strategic effort to reform education rather than an isolated administrative change.

Changes in national high-stakes assessments are inherently gradual and deliberate, a necessity borne out of their profound implications for individuals' futures and broader societal expectations. Documents from the period of policy implementation highlight the significant concern and stress experienced by stakeholders – students, parents, teachers, and educational administrators – underscoring their demand for credibility, fairness, and transparency in the assessment process. This is particularly evident in opinion pieces and petitions submitted by students, which is especially noteworthy given that their representatives had played a significant role a decade earlier in shaping the policy objectives aimed at initiating the reform.

The introduction of digital mathematics tests generated widespread concerns and negative reactions, particularly before the first examination day. Despite the use of digital tools in assessments since the 1980s, many stakeholders expressed apprehension about students' ability to write mathematical notation effectively in a test setting. Implementation records indicate that this challenge was addressed by releasing an open-access version of the exam system, which played a pivotal role in alleviating anxiety among students and teachers. This initiative allowed educators to create custom practice tests and provided students with the opportunity to engage directly with the actual exam platform, offering hands-on experience that proved far more effective in building familiarity and confidence than instructional materials or mock tests alone. Furthermore, the release of an intuitive formula editor and allowing screen captures from familiar mathematics software in the answers addressed concerns with writing mathematical notation in the tests.

When exploring new digital tools for assessment, there is often a tendency to envision sophisticated test items and answer methods. In contrast to Finland's more incremental transition from pen-and-paper assessments to a digital format, Ripley (2009) advocates for a more transformative approach, highlighting the potential of the digital medium to assess 21st-century skills more effectively. Although the reform described in this chapter followed a more migrational strategy, opportunities for more innovative and ambitious developments could be considered if they align with the policy goals of the reform and if the systemic circumstances allow. Furthermore, the adoption of advanced technologies, such as adaptive test designs and artificial intelligence, must comply with stringent regulatory frameworks, such as the European Union's Artificial Intelligence Act (2024), which imposes restrictions on AI systems used in educational settings determining access, admission, or assignment to institutions or training programmes (Annex III), such as the Matriculation Examination tests.

The adoption of digital formats in education, particularly in high-stakes assessment, introduces significant opportunities for advancing educational research. The organising of the Matriculation Examination tests digitally generates extensive datasets, including detailed records of test submissions and log data. These datasets provide a rich source of information on candidates' performance and reasoning processes and could enable educational researchers to gain deeper insights into how students utilise mathematical tools and develop essential competencies. Such insights have the potential to inform the design of future curricula and assessment practices, contributing to more effective and evidence-based educational strategies. Consequently, the integration of digital formats not only transforms assessment practices but also has the potential to establish robust foundations for innovative research in education.

6 Conclusion

In conclusion, in an evolving digital world, ensuring that educational assessment practices remain aligned with digital practices in the rest of society is necessary. However, achieving this alignment requires careful consensus-building and sufficient time and other resources for effective implementation.

The digital reform of the Matriculation Examination in Finland, particularly in mathematics, presented significant challenges for all stakeholders involved. By 2019, the full digitalisation of the examination had been achieved, guided by design principles focused on ensuring the reliable delivery of tests under all circumstances. At the same time, the reform aimed to preserve the pedagogical freedom of educators by equipping the examination system with a comprehensive range of mathematical tools.

Managing the expectations and concerns of stakeholders emerged as a critical component of the implementation process. By providing opportunities for students, teachers, and IT administrators to engage with the digital examination system in advance, it was possible to mitigate anxiety and foster familiarity with the new format. Despite these measures, concerns are inevitable in any major reform, as changes to established practices often evoke apprehension regarding fairness, efficacy, and broader implications. Addressing these concerns requires ongoing dialogue, transparency, and adaptability to ensure trust and confidence in the new system.

This case study offers a practical example of how high-stakes assessments can be successfully digitised, providing a reference point for other education systems considering similar innovations. From 2019 to 2024, the 191,900 digital test submissions in mathematics serve as a testament that such reforms can be implemented in high-stakes assessment even if the test items are predominantly open-ended.

Reference list

Act on the Matriculation Examination, 502/2019. (2019). www.finlex.fi/fi/laki/ajantasa/2019/20190502

Artificial Intelligence Act, Regulation (EU) 2024/1689 of the European Parliament and the Council. (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

The Council of Rectors of Finnish Universities. (2018, February 16). *UNIFIn vararehtorikokous on sopinut ylioppilastutkinnon pisteytystyökalun käytöstä* [Press release]. https://unifi.fi/unifin-vararehtorikokous-sopinut-ylioppilastutkinnon-pisteytystyokalun-kaytosta/?vn/2022-08-10715404.html

The Finnish Association of Teachers of Mathematics, Physics, Chemistry and Informatics. (2011). *MAOL ry:n lausunto lukiokoulutuksen kehittämisen toimenpide-ehdotuksia valmistelevan työryhmän ehdotuksista (Registry no. OKM/45/040/2008) [Statement]*. Ministry of Education and Culture Archives.

The Finnish Association of Teachers of Mathematics, Physics, Chemistry and Informatics. (2016). *Vetoomus sähköisen ylioppilaskirjoituksen työkaluista* [Letter]. Issued by the General Assembly on 19 November 2016. https://maol.fi/app/uploads/2019/06/Julkilausuma_19112016.pdf

Finnish Broadcasting Company. (2024, December 5). *Matematiikka | Abitreenit*. Yle. Retrieved December 5, 2024, from https://yle.fi/abitreenit/matematiikka

Finnish National Agency for Education. (2003). *National core curriculum for general upper secondary schools 2003: National core curriculum for general upper secondary education intended for young people (Registry no. OPH/33/011/2003) [Unofficial translation]*. Finnish National Board of Education.

Finnish National Agency for Education. (2016). *National core curriculum for general upper secondary schools 2015: National core curriculum for general upper secondary education intended for young people (Registry no. OPH/60/011/2015) [Unofficial translation]*. Finnish National Board of Education.

Finnish National Agency for Education. (2020). *Opetushallituksen tilinpäätös 2019 [Financial statement]*. www.oph.fi/fi/tietoa-meista/suunnittelu-ja-seuranta

Finnish National Agency for Education. (2025). *Opetushallituksen tilinpäätös 2024 [Financial statement]*. www.oph.fi/fi/tietoa-meista/suunnittelu-ja-seuranta

Government Decree on the General National Objectives of General Upper Secondary Education and the Distribution of Lesson Hours, 955/2002. (2002). www.finlex.fi/fi/laki/alkup/2002/20020955

Heikkonen, E. (2017, November 30). *Registry no. OPH-3112-2017* [Email]. Matriculation Examination Board Archives.

Hellsten, L., & Rahikka, M. (2018). *Digital answers to a prior mathematics test [Illustration]*. In Vikberg (2018, February 19), Digitaalisia esimerkkivastauksia matematiikan koekysymykseen [Blog post]. Matriculation Examination Board Archives.

Hill, M., & Hupe, P. (2022). *Implementing public policy: An introduction to the study of operational governance* (4th ed.). Sage.

Lattu, M. (2014, November 26). *Sähköisen ylioppilaskokeen kehittäminen jatkuu normaalisti* [Blog post]. https://web.archive.org/web/20150417173031/https://digabi.fi/2014/11/sahkoisen-ylioppilaskokeen-kehittaminen-jatkuu-normaalisti/

Lattu, M. (2015, January 30). *Abitti on auki – sinut on kutsuttu pilottiryhmään!* [Blog post]. https://web.archive.org/web/20170312004910/https://digabi.fi/2015/01/abitti-on-auki-sinut-on-kutsuttu-pilottiryhmaan/

Lattu, M., & Vikberg, T. (2018). Monta reittiä maaliin. *Dimensio: Matemaattisluonnontieteellinen aikakauslehti*, *82*(2), 42–44.

Mäntylä, J. (2018, September 3). Sähköinen ylioppilaskoe tuo lukiolaisille lisää paineita – Miksi päättäjät eivät välitä meidän mielipiteestämme? [Opinion piece]. *Helsingin Sanomat*. www.hs.fi/paivanlehti/03092018/art-2000005813116.html

Matriculation Examination Board. (1981). *Minutes of the matriculation examination board meeting on 2/11/1981 at 2 PM: Section 21 § (133 §)*. Matriculation Examination Board Archives.

Matriculation Examination Board. (2011a). *Lausunto lukiokoulutuksen kehittämisen toimenpide-ehdotuksia valmistelevan työryhmän ehdotuksista (Registry no. OKM/45/040/2008) [Statement]*. Ministry of Education and Culture Archives.

Matriculation Examination Board. (2011b). *Matematiikan kokeen määräykset [Regulations]*. Matriculation Examination Board Archives.

Matriculation Examination Board. (2013a). *Minutes of the matriculation examination board meeting on 3/5/2013 at 1 PM: Section 7 §*. Matriculation Examination Board Archives.

Matriculation Examination Board. (2013b). *Minutes of the matriculation examination board meeting on 20/9/2013 at 1 PM: Section 6 §*. Matriculation Examination Board Archives.

Matriculation Examination Board. (2016, November 28). *Matematiikan jaoksen vastaus MAOL ry:n vetoomukseen* [Letter]. Matriculation Examination Board Archives.

Matriculation Examination Board. (2017a, March 30). *Svar på enkäten om de digitala MAFYKE proven [Statement]*. Matriculation Examination Board Archives.

Matriculation Examination Board. (2017b, May 2). *Uudessa Abitti-versiossa uudistunut vastauseditori ja lisää laitetukea* [Blog post]. www.abitti.fi/blogi/2017/05/uudessa-abitti-versiossa-uudistunut-vastauseditori-ja-lisaa-laitetukea/

Matriculation Examination Board. (2019a, March 29). *Matematiikan kokeet ensimmäisen kerran digitaalisia* [Press release]. https://web.archive.org/web/20210418025754/www.ylioppilastutkinto.fi/ajankohtaista/tiedotteet/750-matematiikan_kokeet_ensimmaisen_kerran_digitaalisia

Matriculation Examination Board. (2019b). *Matematiikan koe, pitkä oppimäärä 26.3.2019 Hyvän vastauksen piirteitä [Assesment criteria]*. https://info.ylioppilastutkinto.fi/hvp/prel/2019_k_m.pdf

Matriculation Examination Board. (2022). *Lausuntopyyntö ylioppilastutkinnon matematiikan kokeiden rakenteiden muuttamisesta [Public consultation]*. www.lausuntopalvelu.fi/FI/Proposal/Participation?proposalId=1616a9d4-fd31-4517-a827-0edf83269198

Matriculation Examination Board. (2024). *Hyvän vastauksen piirteet: FI – Matematiikka, pitkä oppimäärä 26.9.2024 [Assesment criteria]*. https://tiedostot.ylioppilastutkinto.fi/kokeet/2024-09-26_M_fi/grading-instructions.html

Matriculation Examination Board. (2025, January 31). *Oppilaitoskohtaisia tunnuslukuja*, *D4001* [Data set]. Matriculation Examination Board. Retrieved January 31, 2025, from www.ylioppilastutkinto.fi/fi/tietopalvelut/tilastot/oppilaitoskohtaisia-tunnuslukuja

Matriculation Examination Board, & Lokki, J. (Chair). (2013). *Tieto- ja viestintäteknii-kan käyttö ylioppilastutkinnon suorittamisessa: Esiselvitys [Report]*. Matriculation Examination Board Archives.
Ministry of Education. (2008). *Koulutus ja tutkimus 2007–2012: Kehittämissuunnitelma. (Opetusministeriön julkaisuja 2008:9)*. http://urn.fi/URN:ISBN:978-952-485-507-5
Ministry of Education and Culture, & Pirhonen, E.-R. (Chair). (2010). *Lukiokoulutuksen kehittämisen toimenpide-ehdotuksia valmistelevan työryhmän muistio. (Opetus- ja kulttuuriministeriön työryhmämuistioita ja selvityksiä 2010:14)*. http://urn.fi/URN:ISBN:978-952-485-964-6
Parliamentary Ombudsman. (2017, December 13). *Ylioppilastutkintolautakunnan ja Opetushallituksen menettely matematiikan sähköisen ylioppilaskokeen valmiste-lussa ja valmistelun etenemisestä tiedottamisessa (Registry no. EOAK/5991/2017)*. Archive of Parliamentary Ombudsman.
Pekkala, E. (2017, October 5). *Matemaattisten aineiden ylioppilaskirjoitusten on pysyttävä paperisina* [Petition]. Adressit.com. www.adressit.com/matemaattisten_aineiden_ylioppilaskirjoitusten_on_pysyttava_paper
Prime Minister's Office. (2011). *Programme of Prime Minister Jyrki Katainen's government. (Government publications 2011:3)*. http://urn.fi/URN:NBN:fi-fe201704256257
Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment* (pp. 92–98). European Commission Joint Research Centre.
Scott, J. (1990). *A matter of record: Documentary sources in social research*. Polity Press.
Seppänen, R. (2018, September 18). Matematiikan ensi kevään kirjoittajien pitäisi saada suorittaa koe paperilla [Opinnion piece]. *Helsingin Sanomat*. www.hs.fi/paivanlehti/18092018/art-2000005831092.html
Tolvanen, S. (2017, November 27). *Lukion Suuri Adressi (Registry no. OPH-3112-2017)* [Petition]. Matriculation Examination Board Archives.
The Trade Union of Education in Finland. (2011). *OAJ:n Lausunto lukiokoulutuksen kehittämisen toimenpide-ehdotuksia valmistelevan työryhmän ehdotuksista (Registry no. OKM/45/040/2008) [Statement]*. Ministry of Education and Culture Archives.
Union of Upper Secondary School Students. (2007). *Tavoiteohjelma 2008*. Approved at the General Assembly in Varkaus on 24 November 2007. https://web.archive.org/web/20085l8074256/www.lukio.fi:80/service.cntum?pageId=113990
Union of Upper Secondary School Students. (2008). *Tavoiteohjelma 2009*. Approved at the General Assembly in Seinäjoki on 21–23 November 2008. https://web.archive.org/web/20090821104556/www.lukio.fi:80/service.cntum?pageId=113990
Union of Upper Secondary School Students. (2009). *Tavoiteohjelma 2010*. Approved at the General Assembly in Lappeenranta on 21–23 November 2009. https://web.archive.org/web/20100910012903/www.lukio.fi:80/service.cntum?pageId=113990
Union of Upper Secondary School Students. (2010). *Suomen Lukiolaisten Liiton Hallitusohjelmatavoitteet 2011–2015*. https://web.archive.org/web/20100908054453/www.lukio.fi:80/service.cntum?pageId=1115193
Union of Upper Secondary School Students. (2011). *Lausunto lukiokoulutuksen kehittämisen toimenpide-ehdotuksia valmistelevan työryhmän ehdotuksista (Registry no. OKM/45/040/2008) [Statement]*. Ministry of Education and Culture Archives.
Väisälä, O. (2018, September 20). Miksi ihmeessä matematiikan ylioppilaskoe sähköistetään? [Opinnion piece]. *Helsingin Sanomat*. www.hs.fi/paivanlehti/20092018/art-2000005833635.html
Valtavaara, M. (2017, October 5). Matematiikan yo-kokeiden sähköistäminen hämmentää kouluissa – "Kaikki tapahtuu hitaammin kuin ennen: ensin paperille,

sitten koneelle", ihmettelee lukiolainen. *Helsingin Sanomat*. www.hs.fi/suomi/art-2000005395337.html

Vikberg, T. (2017). Valmistautuminen digitaaliseen matematiikan ylioppilaskokeeseen. *Dimensio: Matemaattis-luonnontieteellinen aikakauslehti*, *3*, 7–9.

Vikberg, T. (2018, February 19) *Digitaalisia esimerkkivastauksia matematiikan koekysymykseen* [Blog post]. Matriculation Examination Board Archives.

5

TEACHERS' INSIGHTS FROM DIGITAL FORMATIVE ASSESSMENT COMPARED TO TRADITIONAL PENCIL-AND-PAPER ASSESSMENTS IN ELEMENTARY SCHOOL GEOMETRY

Hassan Ayoob and Shai Olsher

1 Introduction

In the evolving landscape of modern education, the integration of technology has become the cornerstone for advancing teaching methodologies and enhancing student learning. As educational paradigms shift towards student-centred practices, there is a growing need for assessment tools that provide teachers with immediate, actionable insights into students' learning processes. Geometry, a subject inherently reliant on spatial reasoning and visualisation, presents unique challenges in this regard. While widely used, traditional paper-and-pencil assessments offer only a limited snapshot of students' understanding and often fail to capture the dynamic nature of geometric thinking. The static nature of conventional assessments restricts teachers' ability to adapt their instructional strategies to address individual and collective learning needs in the classroom (Black & Wiliam, 1998; Harlen & James, 1997).

Digital formative assessment (DFA) is a promising approach that leverages technology to provide real-time feedback, thereby enhancing the interactive learning experience, particularly in geometry, where understanding is visually and conceptually complex (Bennett, 2011). By implementing DFA within dynamic geometry environments (DGEs), such as GeoGebra, students can engage with geometric figures in an interactive manner by manipulating shapes, observing changes, and exploring properties. This digital approach enables teachers to capture not only the outcomes but also the processes underlying students' reasoning, offering a more comprehensive perspective on student learning than traditional assessments (Leung, 2008).

DOI: 10.4324/9781003533764-5

In this study, a series of DFA activities were designed to enhance teachers' ability to assess and understand student learning in elementary geometry. These DFAs are designed to effectively support the identification, classification, and analysis of student work characteristics for relevant stakeholders, providing educators with detailed insights that can inform instructional adjustments (Ayoob & Olsher, 2023). These activities include example-eliciting tasks (EETs) that encourage students to explore and articulate multiple solutions and offer a nuanced view of their reasoning processes (Olsher et al., 2016). In addition, teachers have access to a dashboard that provides immediate reports at both individual and classroom levels, facilitating efficient formative assessment practices and real-time instructional adjustments. By analysing data gathered from these DFAs, this research aims to determine whether DFA provides teachers with a broader, more detailed view of students' learning, thereby supporting decisions that are responsive to both individual student needs and overall class dynamics.

2 Theoretical background

Formative assessment is an ongoing process that involves the continuous collection and analysis of student data to inform teaching and improve learning outcomes (Black & Wiliam, 2009). This process is deeply embedded in daily interactions between teachers and students, providing real-time feedback processes that enable educators to adapt their instructional strategies to meet the diverse needs of their students (Black et al., 2004). Formative assessment goes beyond mere evaluation; it is integral to the teaching–learning cycle. Effective formative assessment requires teachers to be proficient in collecting and interpreting data on student learning, providing timely and constructive feedback, understanding student learning goals, and adjusting teaching methods accordingly (Heritage, 2007).

Traditional pencil-and-paper assessments, which are widely used in classrooms for both formative and summative purposes, involve students completing written tasks, quizzes, or tests, which teachers then grade manually to evaluate their understanding and progress (Black & Wiliam, 1998). Although these assessments are easy to implement and can be used to assess a broad range of skills and knowledge, they often provide static snapshots of student learning at specific points in time. This static nature can limit the scope of learning, fail to engage students in deeper learning activities, and provide limited opportunities for instructional adjustments (Harlen & James, 1997), thus reducing their effectiveness in guiding teaching and learning (Heritage, 2007).

Visual learning plays a pivotal role in the understanding of geometric concepts in elementary education. Geometry, by its nature, is a visual subject that requires students to grasp the spatial relationships and properties of shapes through diagrams and visual representations (Leung, 2008). The dual approach of using concrete manipulatives and abstract representations, as advocated by Dienes and Golding (1971), helps students develop a deeper understanding of geometric principles. However, traditional pencil-and-paper methods can limit the extent to which students engage with the visual and dynamic aspects of geometry. Static diagrams may not fully capture the transformational nature of geometric figures, which can impede students' ability to dynamically manipulate and understand shapes (Millar, 2017). The advent of technology has transformed geometry teaching and learning significantly. One advantage of DGEs over traditional pencil-and-paper methods is their ability to inform exploration and provide different examples with each manipulation. Each drag or transformation in a DGE offers a new visual example, enabling students to see dynamically a wide variety of instances of a geometric concept. This continuous provision of different visual examples fosters a more comprehensive understanding of the geometric relationships that static methods cannot provide (Butler et al., 2010; Leung, 2008).

Among these technological advancements, digital formative assessment (DFA) leverages technology to provide immediate, specific feedback and facilitates interactive learning experiences, which are key to understanding complex subjects, such as geometry (Bennett, 2011). By integrating DFA within dynamic geometry environments (DGEs) such as GeoGebra, students can dynamically interact with geometric figures, manipulate shapes, observe changes, and explore geometric properties through real-time visual feedback processes (Leung, 2008). The suggested DFA activities are designed to effectively support the identification, classification, and analysis of student work characteristics for relevant stakeholders (Ayoob & Olsher, 2023). These activities include rich tasks with an infinite number of correct solutions structured as example-eliciting tasks (EETs) (Olsher et al., 2016). These tasks feature interactive feedback that highlights students' exploration and reasoning beyond correct or incorrect answers. This design promotes interaction with geometric concepts, enabling students to exemplify, articulate, and reflect on their thought processes, thereby facilitating a critical examination of their reasoning strategies (Stacey & Wiliam, 2013). This type of task involves student-centred assessment, which provides teachers and students with detailed insights into each student's work (Olsher, 2022).

3 Methods

Our goal in this study is to focus on how the use of digital formative assessment tasks changes teachers' ability to assess individual students and the entire class in elementary school geometry in comparison to traditional assessment. To achieve this goal, this study sought to address the following question: How does the use of digital formative assessment tasks change teachers' ability to assess individual students and the entire class in elementary school geometry, in comparison to traditional assessment approaches?

To gain a comprehensive understanding of how digital formative assessments (DFA) impact teachers' ability to assess students in elementary school geometry, a mixed-methods approach was employed. This approach integrates both quantitative and qualitative data to provide a more nuanced analysis of research questions. We analysed the responses gathered through semi-structured interviews. We used open coding (Strauss & Corbin, 2004) to systematically examine teachers' insights in order to identify and categorise the metrics they provide. We aimed to address our research question by delineating the emerging patterns and themes. Additionally, quantitative data were derived from a structured questionnaire administered to teachers following each assessment activity. Specifically, five questions directly relevant to the overarching research question were selected for detailed analysis. This approach enabled us to uncover rich, context-specific information regarding teachers' perspectives on the effectiveness and alignment of DFAs with curriculum goals and students' skills.

3.1 Participants

The participants were nine mathematics teachers from five Arabic-speaking elementary schools in Northern Israel. All schools teach according to Israel's national curriculum. The participants were a diverse group of teachers with varying educational backgrounds and levels of experience. The majority held master's degrees, with teaching experience ranging from 8 to 25 years, in the third to sixth grades.

3.2 Research tools

To examine the metrics according to which elementary mathematics teachers evaluate DFA activities as appropriate for students and curriculum content, we used a teacher questionnaire and DFA units designed in the STEP[1] system (Olsher et al., 2016). The STEP (Seeing the Entire Picture) platform automatically analyses and categorises students' responses to a given task, making the characteristics of students' answers more accessible for teachers (Yerushalmy et al., 2017).

The 12 activities were designed according to the design principles set out in previous research (Ayoob & Olsher, 2023). The activities align with the textbooks and the recommended instruction and distribution of the topic to teaching hours according to the curriculum. The tasks were designed in STEP, and the automatically assessed characteristics of student submissions were designed according to previous literature and the known challenges for each topic.

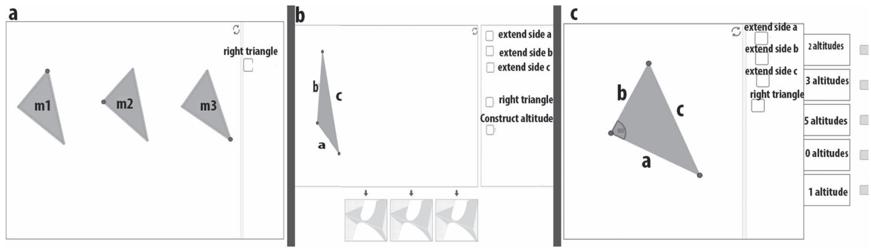


FIGURE 5.1 Three tasks of the right triangle altitude activity: (a) Construct altitudes from the marked vertex, (b) construct three different examples of right triangles and their altitudes, and (c) identify and construct altitudes that coincide with the sides.

The mathematical topics in this study were triangle altitude and polygon area. The activity of the altitudes of the right triangle is shown in Figure 5.1.

The first task (Figure 5.1a) involved three instances of the same right triangle. In each instance, the student was asked to drag the bold vertex to construct an altitude from that vertex and submit it. The triangles are static, and the students can construct segments only from bold vertices. In the second task (Figure 5.1b), the students were asked to construct three different examples of right triangles and then construct one altitude by dragging a point from a vertex. Students can drag any of the vertices of a triangle to create triangles of their choice. In the third task (Figure 5.1c), the students were asked to identify and specify the number of altitudes that coincided with the sides of the triangle that could be drawn. Additionally, students were required to manipulate these altitudes by dragging points from the vertices to 'stretch' the segments, thereby constructing the altitudes. On the left side of each task, the students had the option to select and use several tools, such as displaying the lines extending the sides or using a right-angled triangle ruler.

The questionnaire consisted of 12 questions. For this study, we focused on five questions that were directly relevant to the overarching research question. These questions were selected for a detailed analysis and included Likert scales ranging from 1 to 5. Teachers were asked to assess each DFA in terms of (1) activity difficulty (1 = very easy to 5 = very difficult), (2) curriculum alignment (1 = not aligned to 5 = very aligned), (3) suitability for assessment (1 = not suitable to 5 = very suitable), (4) impact on future instruction planning (1 = no impact to 5 = high impact), and (5) student responses compared to teachers' expectations (1 = unexpected to 5 = very expected). In addition, averages were computed for each question based on ratings.

Finally, semi-structured interviews were conducted with participating teachers to explore their insights and experiences with DFA. The interviews were designed to gather qualitative data on teachers' perceptions, the

TABLE 5.1 Characteristics of the research participants and participation in research tools

Teacher	*Education*	*Seniority*	*Activities implemented*	*Answering the questionnaire*	*Interview*
Teacher L	MA	18	12	12	2
Teacher S	MA	19	12	6	0
Teacher A	MA	15	11	11	2
Teacher RA	MA	20	10	7	0
Teacher N	MA	18	9	9	2
Teacher K	BA	8	9	9	0
Teacher D	MA	20	9	9	2
Teacher B	MA	25	5	5	1
Teacher R	MA	18	1	1	0
		Total	80	69	9

challenges, and the effectiveness of DFA and traditional formative assessments. The interviews began with general questions to create an open and conversational atmosphere, encouraging teachers to share their impressions of the DFA tasks. Subsequent questions focused on exploring their assessment practices, including the tools they commonly used and their approaches to addressing student errors. Teachers were also asked to reflect on how DFA tasks aligned with their expectations of student performance, and whether the tasks supported the curriculum effectively. To gain deeper insights, participants were invited to discuss the specific DFA tasks they had implemented, sharing their observations of student engagement, the challenges faced, and any unexpected outcomes. Interviews were conducted by inviting teachers to share additional reflections or suggestions related to DFA activities.

Table 5.1 provides an overview of the participants and details of their professional and educational characteristics. It includes information on teachers' education level, years of seniority, the grades they teach, the number of activities they have implemented, and their participation in answering the questionnaire and interview. These data serve as a foundation for analysing their involvement and contributions to the study.

3.3 Research setting

The preparatory stages were undertaken before introducing activities to the classroom for both teachers and students on the STEP platform. The first author trained the participating teachers on the STEP platform and its teacher–student interfaces and provided clear guidance on the timing and integration of DFA tasks into their instructional plans. Throughout the academic year, our focus was on ensuring that teachers promptly conducted assessments after teaching each subtopic. STEP recorded all the submissions for each student, and the teachers had access to the students' answers through four different reports provided by STEP. These reports allowed teachers to generate data and analyse submissions according to specific characteristics. STEP also facilitated the anonymous presentation of notable cases and interesting characteristics for discussion purposes, or for providing general or individual feedback in class. After each activity and the ensuing discussions with the students about their submissions, the teachers completed a questionnaire.

3.4 Data collection and analysis

The data for this study were obtained from the responses of nine teachers to five questions in the questionnaire and nine interviews, following the completion of each of the 12 research activities. Responses to the questionnaire were computerised and saved on a spreadsheet. These responses were not used until all the activities were completed. Sixty-nine submissions were collected during the course of this study. Quantitative data analysis focused on calculating the averages, standard deviations, and ranges for each of the selected questions. This analysis provided a statistical overview of teachers' evaluations of DFA activities in terms of difficulty, curriculum alignment, suitability for assessment, impact on future instruction planning, and students' responses to expectations.

Data were collected through nine semi-structured interviews with five teachers. Two interviews were planned with all five participating teachers to capture their immediate and reflective insights into the DFA tasks. However, one teacher was unable to participate in the second interview because of scheduling conflict and personal constraints. Despite this dropout, the data from the first interview with this teacher were retained as they provided valuable insights. The interviews were recorded and saved, and an open coding process was performed after completion of all activities (Strauss & Corbin, 2004). We used an iterative process for qualitative data analysis, which included the following steps: (1) *Preparation*. All interviews were transcribed as accurately as possible, preserving the exact wording and terms used by the teachers. (2) *Initial coding*. The first interview transcript was carefully read to identify and highlight significant sentences, leading to the creation of the initial codes. These codes include references to the DFA, such as the scope of reports, timing, and associated challenges. (3) *Unification of codes*. Similar codes are grouped into meaningful codes. For example, codes such as student reports and teacher reports were unified under the reports code, while those related to corona time distance learning and post-instruction were categorised under the Timing of Implementing code. (4) *Categorisation*. The refined codes were organised under broader general categories to create a structured framework.

This process helped identify key themes and patterns across the interviews. (6) *Validation*. According to Creswell and Clark (2013), validation strategies involve multiple researchers and member checks to enhance the accuracy and credibility of the findings. Approximately 15% of the statements from the interviews were prepared with corresponding categories on two sheets for validation purposes. This validation was conducted by the Mathematics Education Research and Innovation Center (MERI) team, comprising graduate students, curriculum developers, and researchers in mathematics education. In all cases, disagreement led to consensus and, in some cases, modification of the coding scheme.

During some of the interviews, specific data were presented to explain certain characteristics discovered by the teachers, highlighting their surprises. This helped illustrate how teachers utilised DFA reports to understand student performance and adjust their teaching strategies accordingly.

4 Results

The quantitative results presented in Table 5.2 involve evaluating teachers' perceptions of the digital formative assessment tasks used in elementary school geometry. The following five criteria were assessed: activity difficulty, curriculum alignment, suitability for assessment, impact on future instructional planning, and students' responses to expectations.

The results showed a moderate perception of activity difficulty with some variability. The rating statement was: 'On a scale of 1–5, how difficult is the activity, where 1 is very easy and 5 is very difficult?' The curriculum alignment results indicated a strong agreement among teachers regarding the assessment's alignment with the curriculum. Teachers rated their suitability for formative assessment, with a mean score of 4.15, reflecting a positive view of the appropriateness of the assessments for formative purposes. The impact on future instructional planning was rated highly, showing strong consensus on its value. Finally, student responses against expectations had a mean rating of 2.90, the highest standard deviation of 0.97 among the criteria, indicating a wide variation in teacher experiences.

TABLE 5.2 Teachers' ratings on digital formative assessments

Evaluation criteria	*Mean*	*SD*	*MAX*	*MIN*
Activity difficulty	2.59	0.63	4	1
Curriculum alignment	4.54	0.64	5	2
Suitability for assessment	4.15	0.69	5	2
Impact on future instruction planning	4.27	0.55	5	3
Student responses against expectations	2.90	0.97	5	1

4.1 Interview findings

The following section presents the findings of the nine interviews conducted with five of the nine elementary school teachers who participated in this study. Table 5.3 outlines the specific categories and codes referenced by the teachers during the interviews with digital formative assessments (DFA). This table (Table 5.3) shows the number of teachers who mentioned each code. The codes were organised into four categories.

The first category of 'assessment enactment' encompasses how DFA tasks are structured and scheduled within the educational setting. This general category includes two codes: organisation and sequence and timing of implementation.

Organisation and sequence. Teachers consistently highlighted the importance of the organisation and sequencing of DFA tasks. Proper organisation ensures that tasks build on each other logically, helping students to progressively develop their understanding of geometric concepts. For instance, starting with simpler tasks and moving to more complex tasks effectively help scaffold learning. This systematic arrangement supports students in building a strong foundation before they tackle advanced topics. Teacher N emphasised the importance of the sequential organisation of assessment tasks, indicating that each task has its inherent value due to its specific focus. 'Each one [activity] in its time is important because each activity assessed aspects, the first one about perpendicular sides, then about the right triangle; each one has its importance' (Teacher N, interview, 01 February 2021).

Timing of implementing. Teachers pointed out that tasks need to be integrated at appropriate points during instruction to reinforce and assess

TABLE 5.3 Categories referenced by teachers in interviews of DFA

Categories	*Codes*	*Number of teachers that mentioned*
Assessment enactment	Organisation and sequence	5
	Timing of implementation	5
Assessment design	Scope of assessment	5
	Assessment that promotes learning	5
	Static and dynamic tasks	3
	Comparison with traditional assessment	3
Information provided	Reports	3
	Student responses against expectations	4
Interaction and adaptive planning	Classroom discussions	5
	Planning for teaching	5

learning. Immediate implementation after teaching a concept ensures that the assessment is timely and relevant, providing immediate feedback that can be used to adjust instruction. This alignment with the instructional timeline helps maintain the flow of teaching and learning processes. Teacher N expressed appreciation for the comprehensiveness and the digital platform's ability to cover various types of triangles, highlighting the educational potential of such tools: 'The tasks are easy when the student learns the subject . . . having something tangible. . . . It even helps him understand the subject' (Teacher N, interview, 1 February 2021).

The second category of assessment design refers to the various underlying design considerations used to evaluate student learning through the DFA. It consists of four codes: scope of assessment, assessment that promotes learning, static and dynamic tasks, and comparison with traditional assessments.

Scope of assessment. Teachers valued DFA tasks for their broad scope, allowing the assessment of a wide range of topics within geometry. For example, when students were tasked with constructing the altitude of an obtuse triangle, the activity not only assessed their ability to construct the altitude but also evaluated their ability to identify and construct the appropriate type of triangle. This comprehensive approach ensured that various aspects of the students' geometric understanding were evaluated, ranging from basic definitions to more complex skills. The ability to cover a wide range of content within a single assessment framework was highly appreciated by the teachers, as demonstrated by Teacher L, who described digital tasks as encompassing a variety of triangle types and altitudes, indicating a comprehensive approach to the geometry curriculum. 'I felt that you provided all types of triangles with all the possibilities for altitudes, which could be along the extension on the side outside or inside. This included obtuse, right, and acute triangles, which taught what properties the altitude has and the conditions that must be met' (Teacher L, interview, 24 January 2021).

Assessment promotes learning. Teachers mentioned that DFA tasks not only assess but also promote learning, stating that these tasks encourage students to engage deeply with the content, fostering a better understanding and retention of geometric concepts. As Teacher A stated:

> I noticed that the students, especially those who were average and above, learned new strategies through these tasks. For example, when asked about the triangle and requested a similar area, some came up with the strategy of giving an altitude parallel to the base.
>
> *(Teacher A, interview, 11 September 2021)*

This dual role of assessment and learning promotion is often described as a part of authentic assessment practices.

Static and dynamic tasks. Teachers noted that the inclusion of both static and dynamic tasks was valuable for assessment. Static tasks involving fixed shapes (e.g. the predefined triangles in Figure 5.1(a)) provide a clear and consistent context for evaluating procedural skills. By contrast, dynamic tasks (e.g. Figure 5.1(b)) allow students to manipulate geometric figures, promote exploration, and offer deeper insights into their understanding of geometric properties and relationships. Together, these approaches balance accuracy assessment with creative reasoning. As stated by Teacher L:

> You can give the student a more comprehensive view of the subject. For instance, we sometimes draw all altitudes on the board; however, when the student starts moving the vertices around, they begin to see how things connect. It gives them a more comprehensive view than just looking at the board or at the paper.
>
> *(Teacher L, interview, 24 January 2021)*

This variety caters to different learning styles and helps develop a more rounded understanding of subject matter. Traditional assessments often lack this dynamic component, limiting their ability to assess students' interactive and exploratory learning processes.

Comparison with traditional assessments. Some teachers appreciated the ability to directly compare the DFA structure, results, or other aspects with those of traditional assessments. For example, the time required for traditional assessment practices, particularly during the grading phase, results analysis, and extraction of insights. Teacher N pointed out the extensive time she spent on these tasks: 'Analysis takes time with me. I am slow. . . . I scrutinise a lot' (Teacher N, interview, 1 February 2021). This highlights the inefficiency of traditional exams in providing timely feedback and in promptly adapting teaching strategies. In contrast, the use of a computerised platform that offers automatic grading and various interactive and static reports can greatly improve teachers' efficiency. This platform enables quicker analysis and generates comprehensive reports, thereby streamlining the assessment process and allowing teachers to adjust their instructional methods promptly based on the immediate feedback.

The third category, information provided, encompasses the mechanisms through which the DFA tasks provide reports and engage students in the learning process. It includes two subcategories: reports and students' responses to expectations.

Reports. Teachers emphasised the importance of reports provided by DFA tasks. Information is provided immediately; it is specific and actionable, allowing students to engage in a timely and meaningful feedback process. It also provides teachers with valuable insight into student performance. This immediate feedback loop is essential to foster an adaptive learning environment.

Teacher B emphasised the educational value of feedback as a means of deepening understanding and reinforcing learning.

> Yes, of course. The child likes this, especially today, when we do an exam: The child wants to see his grade to receive feedback. This approach was also supported. It's good for the child to have feedback; to see how they solved it, it can also help them understand it more clearly.
>
> *(Teacher B, interview, 19 February 2021)*

Student responses against expectations. DFA tasks often reveal unexpected student responses that are particularly valuable for teachers. These responses offer deeper insights into students' thinking processes and errors, highlighting areas that require further instructional attention. This category underscores the diagnostic power of DFA in uncovering hidden aspects of student understanding that are not typically visible through traditional assessments. As Teacher A stated: 'I noticed that some of the very good students consistently made errors, and it helped me to understand their misconceptions better' (Teacher A, interview, 11 September 2021).

The fourth category, 'interaction and adaptive planning', refers to how DFA tasks facilitate interactions between teachers and students and support adaptive instructional planning. This category includes two subcategories: classroom discussion and teaching planning.

Classroom discussions. Teachers highlighted the value of discussions prompted by the DFA tasks. Discussions between students and teachers help clarify misunderstandings, deepen understanding, and promote critical thinking. These are essential components of formative assessments that facilitate a dialogic learning environment. Such interactions are crucial for developing a deeper understanding of geometric concepts and for fostering collaborative learning environments. Teacher A assessed the role of the reports in the facilitation of the discussion: 'Displaying the answers [teacher report] without names and asking students why they think a solution is correct or incorrect led to valuable classroom discussion. This visual and interactive approach helped students to see and understand their errors better' (Teacher A, interview, 11 September 2021).

Planning for teaching. The insights gained from the DFA tasks significantly influenced the teachers' planning. Detailed data on student performance allowed teachers to tailor their instruction to better meet individual and class needs. Teacher N stated:

> When I taught the obtuse triangle altitude, the students would then apply the DFA; some students found the external height difficult to understand, so I had to explain it again. This was due to the activity, which helped me notice the difficulty.
>
> *(Teacher N, interview, 20 September 2021)*

Adaptive planning is crucial for addressing diverse learning styles and ensuring that teaching strategies are responsive to students' needs.

The results show that the integration of quantitative and qualitative data in this study provides a holistic view of how DFA tasks affect teaching and learning. The high ratings for curriculum alignment, suitability for formative assessment, and impact on future instructional planning were supported by qualitative insights that emphasised the value of logical sequencing, timely feedback, and comprehensive coverage of content areas.

5 Summary and discussion

The results demonstrate that teachers' qualitative feedback on the organisation, timing, and dynamic nature of DFA tasks aligns with quantitative ratings, reinforcing the importance of well-structured and timely formative assessments in supporting effective teaching and learning. The detailed and immediate feedback provided by the DFA tasks, as highlighted in the qualitative interviews, corresponded to high ratings for their impact on instructional planning. This demonstrates how DFA tasks enable teachers to adapt their strategies in real time, thus enhancing their responsiveness to student needs.

The moderate mean rating (2.59) for activity difficulty, with responses ranging from 1 (very easy) to 4 (difficult), suggests that while some teachers found the digital tasks challenging, none considered them overly difficult. This highlights the general accessibility of the tasks while acknowledging the variability in teacher perceptions. This finding aligns with that of Webb and Jones (2009), who emphasised that appropriately challenging tasks are essential in formative assessments for engaging students effectively and gauging their understanding accurately. The lack of extreme difficulty ratings further underscores the importance of designing tasks that are adaptable to diverse student abilities and learning paces, ensuring that they remain manageable yet stimulating.

The high mean rating (4.54) for curriculum alignment underscores the consensus among teachers regarding the relevance and integration of DFA within the established curriculum. This finding resonates with research emphasising that effective assessment practices must align closely with curricular objectives to ensure coherence in teaching and learning experiences (Black et al., 2004; Millar, 2017). Such an alignment facilitates the seamless integration of assessment tasks into daily instructional activities, promoting deeper engagement with subject matter. With a mean rating of 4.15, the suitability of these tasks for formative assessment highlights their perceived value in supporting the ongoing learning processes. This is consistent with the framework presented by Heritage (2007), who posited that formative assessments should serve as feedback mechanisms to inform teaching strategies and support individual student growth. The diversity in teachers' opinions, as indicated by the standard deviation, may reflect varying levels of familiarity and comfort with digital tools in formative assessment contexts. These high ratings for curriculum alignment and suitability for formative assessment align with Shute's (2008) findings. Shute emphasises that well-aligned formative assessments enhance student learning outcomes by providing relevant and timely feedback. The teachers in our study appreciated the alignment of DFA tasks with the curriculum, suggesting that these tasks effectively supported the educational goals and standards set for elementary geometry.

The mean rating of 4.27 for informing future instructional planning indicates a strong agreement among teachers on the utility of digital tasks in shaping their teaching strategies. This aligns with Black and Wiliam's (1998) assertion that effective formative assessment hinges on the ability to use assessment data to adapt instruction to meet learners' needs. A high rating underscores the impact of digital assessments in providing actionable insights, enabling teachers to address students' learning progress and challenges more responsively. Finally, the diagnostic power of digital formative assessments (DFAs) is highlighted by the significant variability in student responses against expectations, with a mean rating of 2.90 and a broad standard deviation. This variability underscores the ability of DFA tasks to provide unexpected insights into the students' understanding of geometry. Such assessments challenge the preconceptions of students' capabilities, revealing both hidden strengths and unforeseen challenges. This finding aligns with Leung's (2008) research on the impact of dynamic geometry environments on student learning, and Butler et al.'s (2010) study on the value of digital formative assessments in uncovering students' abilities that are not typically observed through traditional methods.

Qualitative insights from teachers emphasise the importance of the information provided by DFA reports in offering immediate, specific, and actionable feedback. This feedback not only enables students to understand their mistakes and engage more deeply with the subject matter but also provides teachers with deeper insights into student performance, fostering a more adaptive learning environment. These findings align with Hattie and Timperley's (2007) argument that effective feedback must be timely, specific, and actionable to significantly enhance learning. Teachers noted that DFA tasks often uncovered unexpected responses, even among high-achieving students, providing valuable insights into their thinking processes and highlighting areas that require further instructional attention. The discrepancies between expected and observed outcomes further underscore the importance of adaptive teaching strategies guided by detailed feedback offered by digital tasks, which facilitates a deeper understanding of students' needs and enables teachers to tailor their instruction for more effective learning outcomes. In essence, the unexpected results of digital formative assessments act as catalysts for reflective teaching practices, challenging teachers to adapt their perceptions and instructional approaches to meet their students' diverse needs more effectively.

When examining the broader category related to assessment design, it becomes evident that the success of DFA in promoting learning is largely due to its dynamic nature and comprehensive approach to task design. These codes underscore the importance of dynamic tasks within DGEs, which helps explore and develop a deeper understanding of spatial relationships (Butler et al., 2010; Leung, 2008). Such tasks are more effective than traditional methods for promoting learning, particularly in their ability to cover a wide range of topics and integrate them cohesively.

The 'assessment enactment' category encompasses how DFA tasks are structured and scheduled in an educational setting. Teachers reported that the tasks were logically organised and implemented in a timely manner, facilitating seamless integration into the teaching process. This structured approach contrasts with traditional summative assessments, which may lack a logical sequence, making it more difficult for students to build systematically on prior knowledge. This alignment with the instructional timeline helps maintain the flow of teaching and learning processes (Bell & Cowie, 2001). Teachers have highlighted how DFA tasks are designed to build on one another, ensuring that students grasp fundamental concepts before advancing to more complex ones (Pellegrino et al., 2001). Unlike traditional formative assessments, which may lack immediate adaptability, DFA tasks leverage interactive and thought-provoking designs to assess and promote learning simultaneously. This dual role fosters deeper engagement with the material and enhances the learning outcomes. Moreover, the inclusion of both static and dynamic tasks in the DFA system caters to diverse learning styles, providing a more comprehensive understanding of geometric concepts (Black & Wiliam, 1998; Clark-Wilson, 2010; Shepard, 2000).

The category of 'information provided' focuses on the mechanisms through which DFA tasks provide information and facilitate students' engagement in the learning process. This category was crucial in determining whether digital activity was deemed suitable for formative assessment by teachers in our recent study (Ayoob & Olsher, 2024). Teachers emphasised the importance of immediate and specific feedback provided by the DFA tasks. Traditional assessments often provide delayed feedback, which can reduce their effectiveness in supporting immediate learning needs (Nicol & Macfarlane-Dick, 2006). The instant reports offered by the DFA enabled the teachers to make real-time instructional adjustments. This immediate feedback loop is crucial for formative assessments to guide ongoing learning and instructional improvements (Nicol & Macfarlane-Dick, 2006). The teacher reported that the DFA tasks often revealed unexpected student responses, providing valuable insights into their thinking processes. This diagnostic capability is a key advantage of DFA over traditional assessments, which often miss nuanced insights (Wiliam, 2011). Understanding these unexpected responses allows teachers to tailor their instruction to meet individual student needs.

Interaction and adaptive planning refer to how DFA tasks facilitate interactions between teachers and students and support adaptive instructional planning. The detailed insights gained from the DFA tasks enabled teachers to adapt their instructional planning to meet the needs of their students better. This adaptive approach ensures a more personalised and effective teaching strategy (Bennett, 2011).

Broader implications of adopting DFA include fostering a more adaptive and responsive educational environment. By continuously assessing it as part of the learning process and responding to student needs, DFA can contribute to a more personalised learning experience, potentially reducing achievement gaps (Bennett, 2011). This is particularly relevant in diverse classrooms, where students have varying levels of prior knowledge and learning paces. Further research should explore the long-term impact of DFA on student outcomes, particularly in different subject areas and educational levels. Studies have also investigated the effectiveness of specific types of DFA tasks in promoting higher-order thinking and problem-solving abilities (Pashler et al., 2007). The findings of this study indicate that digital formative assessments can significantly enhance teachers' ability to assess and adapt instructions in elementary school geometry. A carefully designed DFA that is aligned with the curriculum and provides timely, detailed feedback and the ability to uncover deeper insights into student understanding positions it as a formative alternative to traditional assessment. Addressing these challenges and leveraging the strengths of DFA can lead to more responsive and personalised education, ultimately improving student outcomes.

This study highlights the utility of DFA tasks in elementary geometry, particularly for dynamic and visual topics, such as triangle altitudes and polygon areas. However, the focus on geometry limits its generalisability to other mathematical domains, such as algebra, where interactive visual elements may be less impactful. Additionally, while teachers rated the tasks favourably, the absence of qualitative data explaining their questionnaire responses limited their ability to interpret their perceptions fully. The structured nature of DFA tasks, while providing immediate feedback and logical progression, may also lack the flexibility of traditional formative assessments to adapt to diverse classroom needs in real time. Future research should explore the applicability of DFA across different subjects and investigate how teacher training influences practical integration.

Note

1 For more details, see https://step.haifa.ac.il/english/.

Reference list

Ayoob, H., & Olsher, S. (2023). Curriculum-aligned digital formative assessment (DFA) for elementary school geometry. In P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, & E. Kónya (Eds.), *Proceedings of the thirteenth congress of the european society for research in mathematics education (CERME13)* (pp. 4093–4101). Alfréd Rényi Institute of Mathematics and ERME. https://hal.archives-ouvertes.fr/hal-04417260

Ayoob, H., & Olsher, S. (2024). Exploring metrics: Elementary mathematics teachers' evaluation of digital geometry assessment activities technology. In P. Iannone, F. Moons, C. Drüke-Noe, E. Geraniou, F. Morselli, K. Klingbeil, M. Veldhuis, & S. Olsher (Eds.), *Proceedings of FAME 1 – feedback & assessment in mathematics education (ETC 14)* (pp. 47–55). Utrecht University and ERME. https://hal.science/hal-04807902v1/file/FAME1_Proceedings.pdf

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536–553. https://doi.org/10.1002/sce.1022

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Black, P., Harrison, C., Marshall, L. B. & William, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21. https://doi.org/10.1177/003172170408600105

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Butler, D., Jackiw, N., Laborde, J. M., Lagrange, J. B., & Yerushalmy, M. (2010). Design for transformative practices. In C. Hoyles & J.-B. Lagrange (Eds.), *Mathematics education and technology-Rethinking the terrain* (pp. 425–437). Springer.

Clark-Wilson, A. (2010). *How does a multi-representational mathematical ICT tool mediate teachers' mathematical and pedagogical knowledge concerning variance and invariance* [Doctoral dissertation, Institute of Education, University of London].

Creswell, J. W., & Clark, V. L. P. (2013). *Designing and conducting mixed methods research* (3rd ed.). Sage Publications, Inc.

Dienes, Z. P., & Golding, E. W. (1971). *Approach to modern mathematics*. Herder and Herder.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379. https://doi.org/10.1080/0969594970040304

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, *89*(2), 140–145. https://doi.org/10.1177/003172170708900210

Leung, A. (2008). Dragging in a dynamic geometry environment through the lens of variation. *International Journal of Computers for Mathematical Learning*, *13*, 135–157. https://doi.org/10.1007/s10758-008-9130-x

Millar, R. (2017). Using assessment materials to stimulate improvements in teaching and learning. In K. Hahl, K. Juuti, J. Lampiselkä, A. Uitto, & J. Lavonen (Eds.), *Cognitive and affective aspects in science education research* (pp. 31–40). Springer. https://doi.org/10.1007/978-3-319-58685-4_3

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Olsher, S. (2022). Teaching to collaborate: Automatic assessment-based grouping recommendations for teaching. In U. T. Jankvist, R. Elicer, A. Clark-Wilson, H.-G. Weigand, & M. Thomsen (Eds.), *Proceedings of the 15th international conference on technology in mathematics teaching (ICTMT 15)* (pp. 214–223). Danish School of Education, Aarhus University. https://doi.org/10.7146/aul.452

Olsher, S., Yerushalmy, M., & Chazan, D. (2016). How might the use of technology in formative assessment support changes in mathematics teaching? *For the Learning of Mathematics*, *36*(3), 11–18. https://doi.org/10.2307/44382716

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193. https://doi.org/10.3758/BF03194050

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. https://doi.org/10.17226/10019

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (Vol. 27, pp. 721–751). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Strauss, A. L., & Corbin, J. (2004). Open coding. In A. Bryman (Ed.), *Social research methods: A reader* (pp. 303–306). Oxford University Press.

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *16*(2), 165–184. https://doi.org/10.1080/09695940903075925

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Yerushalmy, M., Nagari Haddif, G., & Olsher, S. (2017). How might the design of tasks for online formative assessment support our understanding of students' conceptions? *ZDM – The International Journal on Mathematics Education*, *49*(5), 701–716. https://doi.org/10.1007/s11858-017-0871-7

6

ADAPTIVE FEEDBACK AND (DIGITAL) ASSESSMENT OF CONCEPTUAL KNOWLEDGE

Guido Pinkernell and Hans-Georg Weigand

1 Introduction

Digital assessment offers new types of assessment with interactive tasks, automatic assessment, and adaptive feedback. These new types involve new tasks and assessment problems on the one hand, but also new ways, styles, or strategies for solving these problems on the other (see, for example, Drijvers & Sinclair, 2023). Interactive software, such as computer algebra systems (CAS), dynamic geometry systems (DGS), and spreadsheets, provide opportunities for interactive and dynamic tasks that are not possible with pen and paper. Adaptive feedback enables the individualisation of learning. However, problems and difficulties with digital assessment must also be taken into account. It is easier to assess calculation and algorithmic tasks than tasks that focus on understanding of concepts. In this article, we focus on the latter. We start with the distinction between procedural and conceptual knowledge, explain the basics of adaptive feedback in formative assessment, and give examples from the area of functions and equations to illustrate how tasks might be developed and show different types of feedback in the solution process of these tasks.

2 Understanding of mathematical concepts

Understanding mathematical concepts is a complex process, and the development of understanding is a long-term process. There are many ways and models to describe the concept of *understanding*. Skemp (1976, 1986) distinguishes between instrumental and relational understanding. He describes *instrumental understanding* as the ability to solve problems by applying memorised rules without knowing their reasons, for example, solving equations

DOI: 10.4324/9781003533764-6

or calculating the derivatives of a function. *Relational understanding* is the ability to solve problems correctly and to explain the reasons for their doing. A more detailed way of classifying understanding is to describe the concept at different levels of cognitive development (Vollrath, 1984; Greefrath et al., 2016).

- At the level of *intuitive understanding*, learners should give simple examples of a concept and use different representations.
- At the level of *content-related understanding*, the focus is on knowing and recognising properties of the concept.
- The level of *integrated understanding* makes it possible to show relationships between the concept and other concepts, as well as relationships between properties.
- The level of *critical understanding* is characterised by the specification of formal definitions and by reasoning and argumentation.

To determine learners' level of understanding, their answers and statements to questions or their actions in problem-solving processes have to be interpreted.

The development of understanding of concepts is one of the most important goals in mathematics classrooms. Understanding requires knowledge and the ability to use this knowledge for explanations, in showing connections, or in problem-solving situations. With regard to mathematical knowledge, it is a widely accepted approach to distinguish between two types: *conceptual knowledge* and *procedural knowledge* (e.g. Hurrell, 2021).

2.1 Conceptual and procedural knowledge in mathematics teaching and learning

An overview of these two types of knowledge can be found in Rittle-Johnson and Schneider (2015). With regard to their categorisation, they refer to Hiebert and Lefevre (1986):

> Conceptual knowledge is characterized most clearly as knowledge that is rich in relationships. It can be thought of as a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information.
>
> *(p. 3)*

Furthermore, a distinction can be made between conceptual knowledge (CK) about concepts or principles (general principles) and about the principles' underlying procedures (Crooks & Alibali, 2014). CK about a mathematical concept refers to knowledge about the properties of the concept, the relationships between different properties, and the relationships to other concepts.

Having CK of a mathematical procedure means knowing the basics of the procedure and its possible applications, being able to assess the limits of the procedure and see it in relation to other procedures. For example, CK about equations is characterised by knowledge about the classification of equations, the conditions for possible solutions, or the possible number of solutions to an equation.

Procedural knowledge (PK) is characterised by Hiebert and Lefevre (1986):

> Procedural knowledge, as we define it here, is made up of two distinct parts. One part is composed of the formal language, or symbol representation system, of mathematics. The other part consists of the algorithms, or rules, for completing mathematical tasks.
>
> *(p. 6)*

Various empirical studies show that pupils and students have significantly more problems with tasks in the area of CK than with tasks in the area of PK (Barumbun & Kharisma, 2022; Bressoud et al., 2017; Zuya, 2017). However, tasks cannot always be clearly characterised as requiring either conceptual or procedural knowledge, as both types are interlinked and often are both necessary to complete a task (Haapasalo & Kadijevich, 2000; Hechter et al., 2022). This is also expressed in the construct 'procept', which refers to the omnipresent duality of object and process:

> We define a procept to be the amalgam of process and concept in which process and product is represented by the same symbolism.
>
> *(Gray & Tall, 1994, p. 117)*

Nevertheless, it seems particularly useful to distinguish between these two types of knowledge when analysing processes of knowledge development and application. Conceptual and procedural knowledge are a part of understanding. Understanding a mathematical concept also includes the ability to adequately apply both types of knowledge for solving problems in mathematical and non-mathematical contexts and to establish relationships and differentiations with other concepts.

2.2 Basic mental models

A central basis for the development of mathematical concepts, and thus CK, are 'basic mental models' (vom Hofe & Blum, 2016). A *basic mental model* (BMM) of a mathematical concept is a content-related interpretation that gives meaning to this concept (vom Hofe & Blum, 2016). The concept of BMM can be used both in a *normative* (prescriptive) and an *individual* (descriptive) sense (see Weigand et al., 2017):

- *Normative* BMMs are the answer to the question 'How should students generally and *ideally* think of a given mathematical concept?' They are identified by didactical analyses of the mathematical concept. They can be used as educational guidelines and to specify learning objectives for mathematics lessons. The determination of BMMs is a didactical challenge for researchers and requires a subject-oriented classification of mathematical and real-life situations of the concept.
- *Individual* BMMs are individual mental models or concepts students *actually* develop in learning processes and problem-solving situations. They can vary from or represent only part of normative BMMs – they even can be based on misconceptions. Individual BMMs are the result of the personal development of meaning and the integration of the concept into an individual's personal worldview.

BMMs of mathematical concepts can be considered within the theoretical framework of 'concept image – concept definition' (Tall & Vinner, 1981). They are parts or subsets of the 'concept image' of a mathematical concept. While *concept image* refers to all individual mental images identified with the concept, BMMs – especially when corresponding to normative BMMs – are the *core* or *central components* of these images.

For example, BMMs of equations are based on mathematical aspects of equations: the definition of an equation, the use of the equals sign, and the relationship between the equation concept and the concepts of algebraic expressions and functions. Four BMMs can be distinguished (Weigand et al., 2022):

- *Operational BMM*. The equals sign is seen as an operational sign, which indicates a reading direction of the equation in the sense of a 'resulting-in' sign.
- *Relational BMM*. The equals sign is seen as a relation sign. The variable here is understood as an unknown which has to be determined.
- *Functional BMM*. An equation $T_1(x) = T_2(x)$ is a comparison of two expressions which are understood as functions with $y = T_1(x)$ and $y = T_2(x)$.
- *Object BMM*. An equation is regarded as a mathematical object that has characteristic properties, such as the number of possible solutions, the definition range, or special solution algorithms.

3 Formative assessment and feedback

3.1 Summative and formative assessment

Assessment is an important step in learning. It asks the learners to step back from the process of learning and gives proof of their state of learning.

Especially when assessment is followed by feedback, it can initiate new learning based on an analysis of the assessment results (Murphy et al., 2023). While in large-scale settings the instruments for assessment are mostly written surveys, in classroom learning, various forms of individual and social assessment are described, among those group, peer, and self-assessment (Suurtamm et al., 2016).

Following a widely accepted distinction (cp. Black & Wiliam, 1998), assessment serves two main purposes: summative and formative. *Formative assessment* is distinguished from other purposes of assessment by its effect on the learner. Essentially, places and people, also methods and instruments, could be anything that fosters learning: '[Formative assessment] is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning' (Klenowski, 2009; also Black & Wiliam, 1998, 2009).

3.2 Tasks and feedback in formative assessment

As described earlier, the role of tasks in assessment is generally diagnostic. The tasks must 'reflect the mathematics that is important to learn and the mathematics that is valued' (Suurtamm et al., 2016), in all its possible manifestations, namely, content, exercises, and competences. Although there are certainly similarities between tasks in summative or formative assessment, the tasks also differ depending on the form of assessment in which they are used: While tasks for larger-scale summative psychometric investigations must follow requirements of local independence, no ambiguity, and unidimensionality, tasks for a contribution to learner progress can be ambiguous, stimulate reflection, initiate discussion, address the subject matter from different angles (i.e. cover an area of knowledge interdependently), and provide insights for further learning (Suurtamm et al., 2016).

Feedback in educational settings is generally conceptualised as information about a learner's performance or understanding in a given field of knowledge (Hattie & Timperley, 2007). In formative settings, feedback informs learners not only about 'how they stand' but also about 'what to do next' (Stacey & Wiliam, 2013; Kluger & DeNisi, 1996). For feedback to guide learning, it is therefore necessary 'to gather and analyze information about the progress of students, with the intention of improving instruction in real time' (Moreno & Pineda, 2020). To be more precise, it is the feedback that makes formative assessment an integral part of a learning process.

4 Assessing CK and giving feedback that supports CK

It is generally easier to test and assess PK than CK (e.g. Barumbun & Kharisma, 2022; Bressoud et al., 2017). Assessing understanding of mathematical concepts and the adequate application of mathematical concepts in intra- and extra-mathematical problem situations cannot be done in an algorithmic way but must be considered within the framework of a model of understanding. For this reason, Hoogland and Tout (2018) also point out the 'risk of focusing too much on assessment of lower order goals, such as the reproduction of procedural, calculation based, knowledge and skills' (p. 675).

Assessing CK requires, on the one hand, content that should be assessed and which goals should be achieved with this content. On the other hand, there is a need for tasks whose solutions should show the extent to which CK is present.

4.1 Task design: selecting tasks that address aspects of CK

CK and PK cannot be separated from each other. For this reason, tasks relating to CK must always be seen in relation to PK. For example, understanding quadratic equations requires knowledge of the number of solutions as depending on the equation (resp. the discriminant), but also the ability to explicitly calculate the solutions. Another example is the understanding of algorithms, which means knowing the meaning and structure of an algorithm, but also being able to apply this algorithm when required.

Tasks are one of or perhaps even the decisive element in learning mathematics (Watson & Ohtani, 2021). When developing tasks, four aspects in particular need to be considered (e.g. Suurtamm et al., 2016):

- The presentation of the task
- The way in which learners are asked to complete the tasks
- The evaluation of learners' answers
- The feedback given to learners

With regard to the assessment and development of CK, it is important that the solution process, that is, the way in which the problem is solved, and thus process skills, such as presenting, reproducing, recognising, describing, or justifying and arguing, are always included in the assessment. There is a big variety of assessment tasks. There might be, for example, portfolio tasks, multiple-choice tests with well-validated CK answers, or questions with open answers.

4.2 Feedback design: addressing aspects of CK

The first decision in formative feedback design concerns its content: What needs to be included to serve the intended learning effect? Next to mere information about correctness and performance, there are the many forms of elaborated feedback that open up various possibilities for a differentiated feedback design. Narciss (2004) identifies possibilities, including explanations and hints addressing, among others, mistakes, procedures, and concepts. Building on these, we will take on a decidedly mathematics educational perspective when we look at feedback content for fostering conceptual knowledge.

Content is shaped by the role we attribute to the learners in feedback situations. From a constructivist perspective, we assume that feedback needs to encourage the learner to actively engage with the object of learning. 'Feedback then becomes not a control mechanism designed by others to corral the learner, albeit in desirable ways, but a process used by learners to facilitate their own learning' (Boud & Molloy, 2013).

Content is also shaped by the function we attribute to formative feedback. Following Hattie and Timperley (2007), functions could encompass the fostering of 'surface understanding' or 'deep understanding'. In this chapter, the function of feedback is to foster conceptual knowledge of mathematical objects. Hence, feedback content should derive from the ideas of building conceptual knowledge which are outlined earlier in Section 1.2.

Considering all this, we suggest the following three parameters for guiding through the process of deciding on formative feedback (cf. Pinkernell, 2024):

- *Addressing the learning object.* Should the object be addressed with the aim of optimising performance for the given task, or should a basis of understanding be laid for this and thematically related requirements? This design parameter differentiates in terms of task and process level in the model of Hattie and Timperley (2007), assuming that suitable foundations of understanding, such as BMMs, can prove to be viable beyond local application in the given task.
- *Adapting to learner characteristics.* Should every learner receive the same feedback for every answer, or should the feedback be differentiated in terms of answer characteristics and other learner characteristics? This design parameter follows the differentiation of cognitive and other learner characteristics, for example, well-researched systematic errors or misconceptions, and also levels of proficiency or understanding.
- *Activating the learning action.* Should the feedback provide all the necessary information about what is needed to master the given task, or should it initiate an independent (re)construction of knowledge? In the first case, the feedback is presented for being received passively; in the second case, the feedback encourages to actively engage with new learning material.

5 Transformative digital tasks and feedback design

The use of digital tools while working with tasks allows more experimental, heuristic work when solving problems and offers opportunities for realistic modelling tasks (Fahlgren et al., 2021). However, this requires time to try out different approaches and pursue them in sufficient depth. If summative assessment is carried out in the context of limited tests, where the usual examinations in the form of written tasks predominate, then such ways of working are difficult to achieve. The question of the use of digital media in examinations must therefore also go hand in hand with a discussion of alternative forms of examination, such as portfolios, individual work. and project work (Ball et al., 2018).

The use of digital media raises a number of questions about task design: What new types of questions are possible with digitally supported tasks, for example, when films, simulations, or interactive applets are integrated? How do the working methods of digital tasks differ from pen-and-paper tests, and where are the similarities, for example, with regard to the necessary prior knowledge or the preparation of sketches on paper (in addition to the digital tool)? How do digitally set examination tasks influence previous or subsequent learning processes?

5.1 Digital task design

In many respects, digital tasks open up new possibilities in terms of setting tasks, posing problems, and new ways of solving these tasks. Especially if we address conceptual understanding in assessment, a focus on its transformative potentials (Ripley, 2009; Hegedus & Moreno-Armella, 2009; Fahlgren et al., 2021) seems most fitting. Here, we concentrate on the *digital representation* and *interaction* of the object of learning as this has, in our view, direct implications for a learner's 'grasp' or understanding. Among the transformative potentials for representing mathematical objects, the following have been listed repeatedly (Pinkernell et al., 2023; Drijvers et al., 2016; Hegedus & Moreno-Armella, 2009):

5.1.1 Dynamisation of representations of mathematical objects

Definitions of mathematical concepts often encompass a large range of examples, which on traditional paper are represented next to each other. A digital representation can represent the definition by a single dynamic construction which, in situations of learning and assessment, can be changed into its many shapes that the definition allows. A single dynamic construction thus can be seen as an appropriate representation of the class of mathematical objects within the range of its definition (Figure 6.1).

5.1.2 Interaction with representations of mathematical objects

Interactivity adds to the dynamisation of representations by allowing learners to actively explore the class of objects defined by its construction. In assessment, interactive elements are the medium with which learners can demonstrate the 'example space' (Watson & Mason, 2005) that they associate with a mathematical concept (Yerushalmy, 2005). Moreover, the transfer between

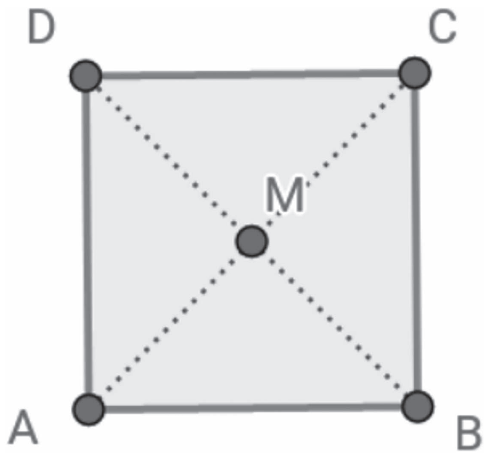


FIGURE 6.1 An interactive GeoGebra applet of a quadrangle constructed as a parallelogram. Its initial shape is that of a quadrangle, but it can be changed into the many shapes that the definition of a parallelogram allows.

Source: Pinkernell et al. (2023).

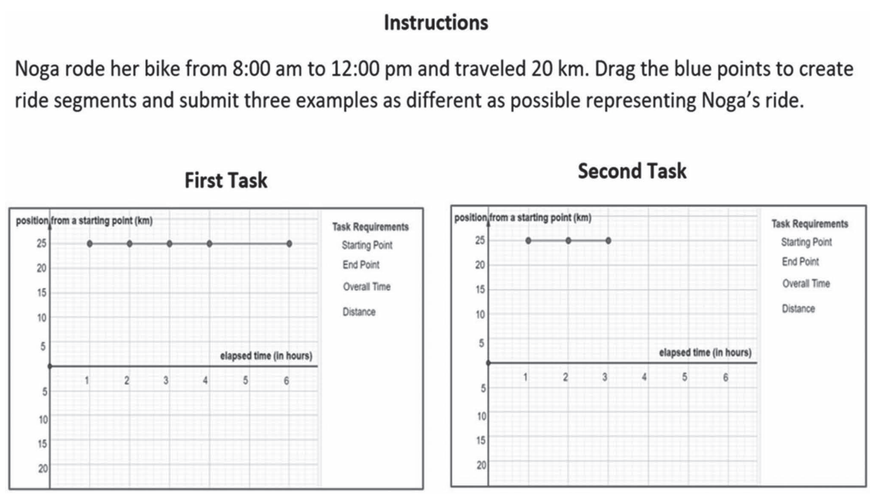


FIGURE 6.2 Three examples are to be provided for each task as representations of the same bike ride. They need to be 'as different as possible', thus widening the space of possible examples.

Source: Yerushalmy et al. (2023).

representations, for example, from pen-and-paper to digital representations, has to be learned and practiced (see also Iannone et al., 2025 in this book).

5.1.3 Multiplicity of representations of mathematical objects

Interconnected multiple representations allow for exploring and demonstrating how syntactically different representations of the same object relate, thus grasping the abstract essence of the mathematical concept (Duval, 1999).

Dynamic, interactive, and multiple representations of mathematical objects seem more appropriate for phases of explorative learning rather than for moments of assessment. In fact, as Stacey and Wiliam (2013) have expressed it, digitisation seems to have a 'blurring' effect on the boundaries between assessment and learning. And so in formative situations where assessment comes with instant feedback for further learning (Fahlgren et al., 2021), dynamic, interactive, and multiple elements show the potential for fostering understanding not only in the tasks but also in the feedback, which both will be outlined in the following sections.

5.2 Digital feedback design

The transformative potential of the digitalisation of learning material can also be used for feedback that is intended to initiate active learning processes in the sense of formative assessment. Following the three design parameters from paragraph 3.2, dynamic, interactive, and multiple representations of mathematical objects can enrich formative feedback. They can open up new aspects of related knowledge and invite learners to actively engage with the object of learning. For the practical implementation, we used the assessment authoring system STACK, which, based on technical contributions from the *AuthOMath* project (Pinkernell et al., 2023), allows the integration of GeoGebra applets in questions and feedback.

5.2.1 Addressing the learning objects

Both aspects of knowledge of a learning object – procedural and conceptual – can be addressed in the feedback. Figure 6.3 describes the change between expression and graph required in the task as a solution procedure to be carried out step by step without addressing the conceptual foundations.

Feedback can also address conceptual basic knowledge instead of procedural activities. This is particularly useful and expedient if fundamental deficits in understanding become visible when solving a task. The feedback in Figure 6.4 shows such a situation: The learner knows transformation algorithms and uses them consistently, but incorrectly. This indicates that the learner has probably not understood the mathematical meaning of the

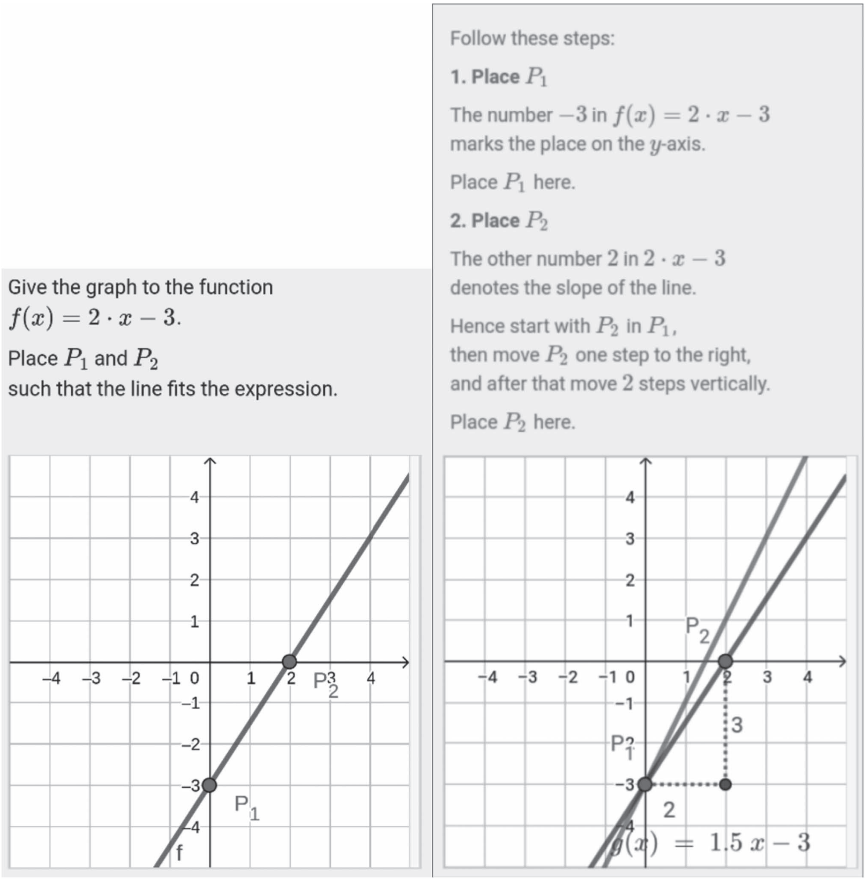


FIGURE 6.3 Changes between expression and graph.

equivalent transformation. This is why the feedback starts with the basics of equivalence transformations. The learner is asked to reconstruct the basics again independently using the explanatory model of the balances.

5.2.2 Adapting to learner characteristics

Non-adaptive feedback gives the same answer for every problem solution. Figure 6.4 is used again as an example. Although the sequence of the balance model shown is adapted to the specific task, it is conceivable as a response to every incorrect solution, and also for correct solutions. This seems to make sense, as solving an equation correctly does not necessarily mean that the learner has the conceptual knowledge for equivalence transformations.

Adaptive feedback, on the other hand, provides specific responses to individual solutions. It reacts, for example, to systematic errors or misconceptions,

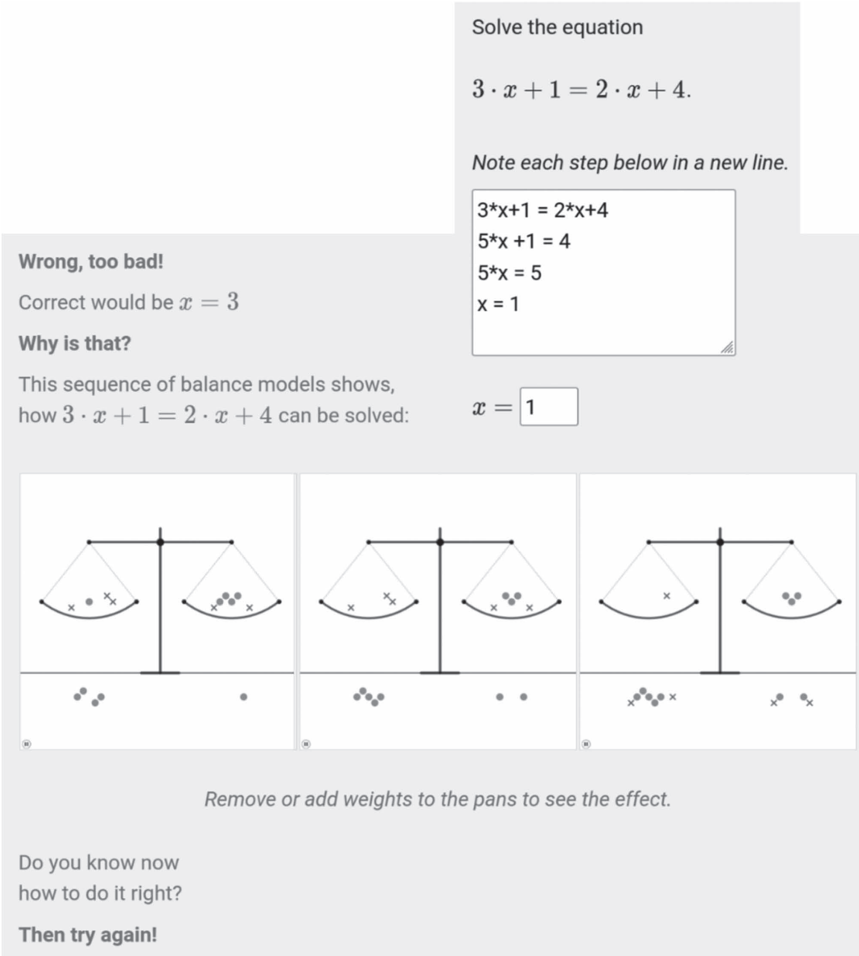


FIGURE 6.4 Feedback addresses basic knowledge of equivalence transformations.

or also to particular correct solutions that make a specific answer appear meaningful. Digital adaptive feedback is based on an automatic analysis of the answer, the mathematical solution of which allows it to be assigned to an answer category. In the first of the following two examples (Figure 6.5), it is assumed that if there is only one sign error, the learner should, in principle, have the necessary knowledge, and therefore, a simple 'nudge' appears to be sufficient.

The task in Figure 6.5 asks to solve linear equations of various kinds, for example, with or without brackets. The feedback then checks the selected (correct) solution strategy as to its efficiency. In order to promote strategic flexibility, reflection on the efficiency of the chosen procedures is encouraged.

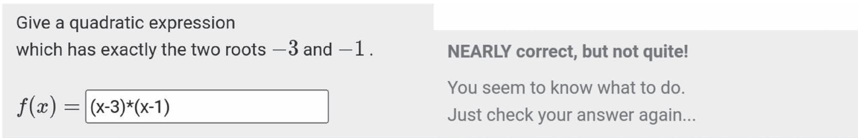


FIGURE 6.5 Feedback as a hint for reconsidering the solution.

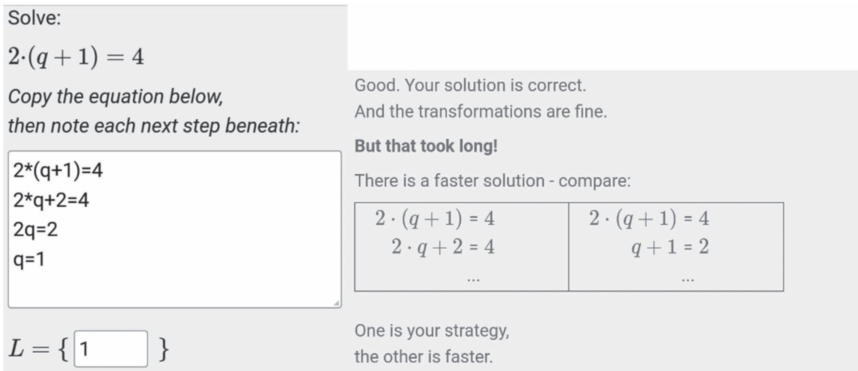


FIGURE 6.6 Adaptive feedback encouraging reflection on the efficiency of the solution.

5.2.3 Activating the learning action

Feedback can support a receptive attitude or encourage active learning. The integration of interactive elements can support the activation of the learner (Barana et al., 2021). The feedback in Figure 6.7 provides all the important information for successfully completing each version of the randomised task. There is no need to reconstruct the necessary knowledge again.

Activating feedback encourages the learner to work on the deficits by themself. Such feedback can contain complete learning tasks in which, for example, interactive elements initiate learning activities that focus on deficits which became apparent in the answer (Figure 6.8).

5.3 Digital task design and digital feedback design

For tasks in real learning, teaching, or assessment situations, elements of digital task design and digital feedback design flow together within the framework of formative assessment. The following tasks are examples of this. They especially include all the aspects of task design and feedback listed in Sections 4.1 and 4.2. We limit ourselves to examples from the field of equations and emphasise the BMMs of equations in particular (see Section 1.2).

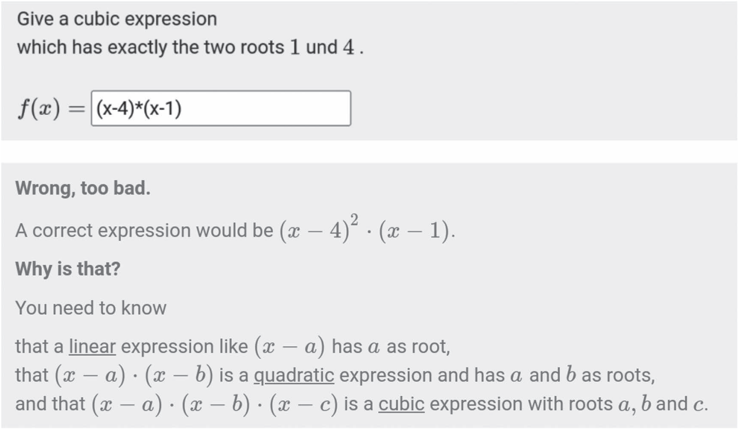


FIGURE 6.7 A randomised task with adaptive feedback.

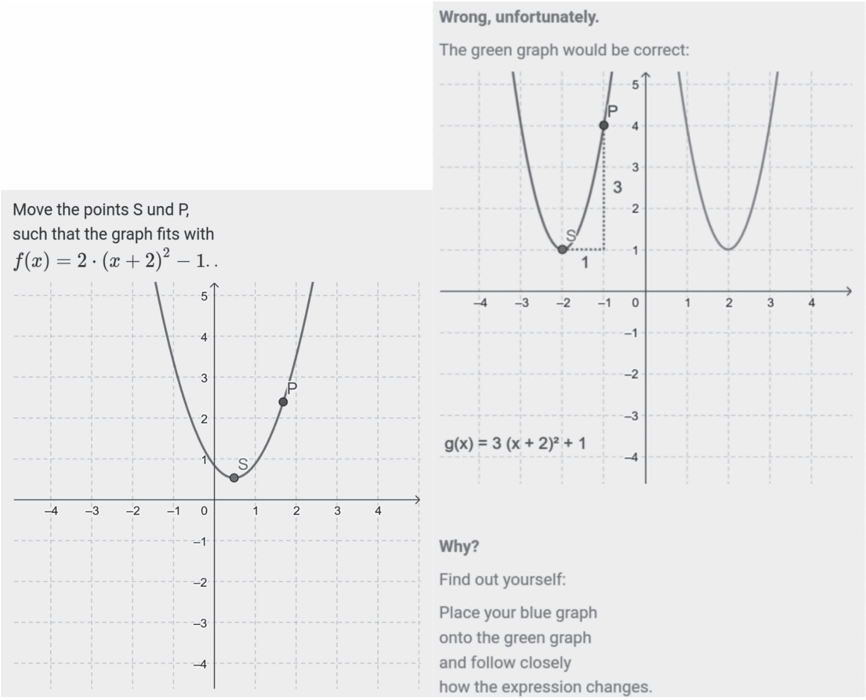


FIGURE 6.8 Feedback with a learning activity.

5.3.1 Parabola inverse: functional BMM

The task shown in Figure 6.9 asks the learner to determine a coordinate system where the given parabola is the graph of a function with the zeros $x_1 = -1$ and $x_2 = 3$.

A coordinate system can now be displayed (Figure 6.9), which can be moved, rotated, and rescaled. The size and the position of the parabola are fixed. The task asks to adjust the coordinate system so that the parabola shows the graph of a function that has the two given zeros. The interactive graph allows experimental approximate solutions.

When adapting the coordinate system, learners might realise that there could be other solutions too (Figure 6.10). The question of the number of different solutions can then be the subject of a discussion in the learning group.

The solution to this problem requires inverse thinking compared to the usual problems involving quadratic functions: The coordinate system for a given graph is not given but must first be found. The difficulty of the problem is further increased by the fact that the parabola is 'oblique' with respect to the usual representation of function graphs.

In many respects, the task requires CK and the activation of the functional BMM of an equation: Firstly, there is the question of the relationship between a function graph or – more generally – a geometric object and its

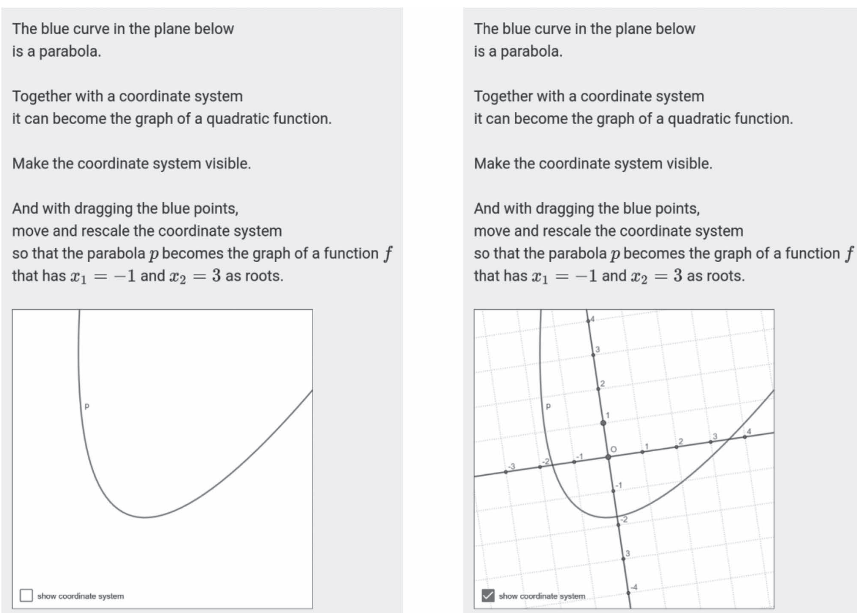


FIGURE 6.9 A parabola in an empty plane (*left*) and with a coordinate system switched on (*right*).

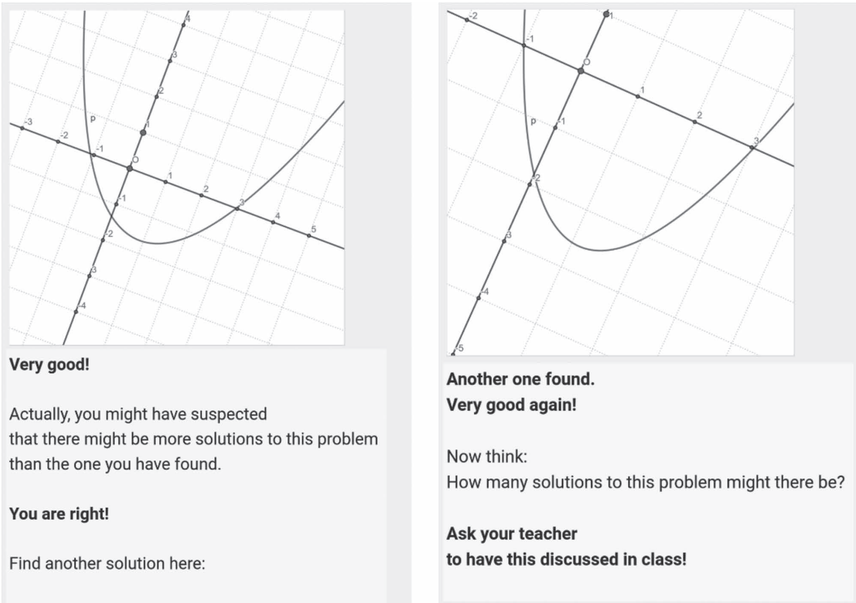


FIGURE 6.10 Two solutions of the problem and adaptive feedback.

representation in a coordinate system. Then there is the significance of symmetry properties of quadratic functions – or the parabola – for the position of the coordinate system. Finally, the question of the relationship between the position of the parabola and the zeros of a quadratic function arises. The feedback challenges learners to reflect on the solution. Concerning the aspects of task design (cf. Section 4.1) and feedback (cf. Section 4.2), the used representations are dynamic and interactive. The multiplicity could be included if you also refer to the numeric and symbolic representations. The adaptive feedback is on the content level; however, it also anticipates learners' actions and expectations.

5.3.2 Exponential and trigonometric equations: relational and functional BMM

How many solutions does the equation $2^x = 1 + \sin(x)$, $x \in \mathbb{R}$, have?

There is no closed solution expression for this equation. The obvious attempt to solve this problem algebraically with GeoGebra's solve operator shows the problem: The solution set is said to be 'undefined' (Figure 6.11).

If the solution on the symbolic level does not lead to the goal, it is always a good strategy to look at the situation graphically. The representation of the equation in terms of the functional BMM (Figure 6.12) shows the two graphs

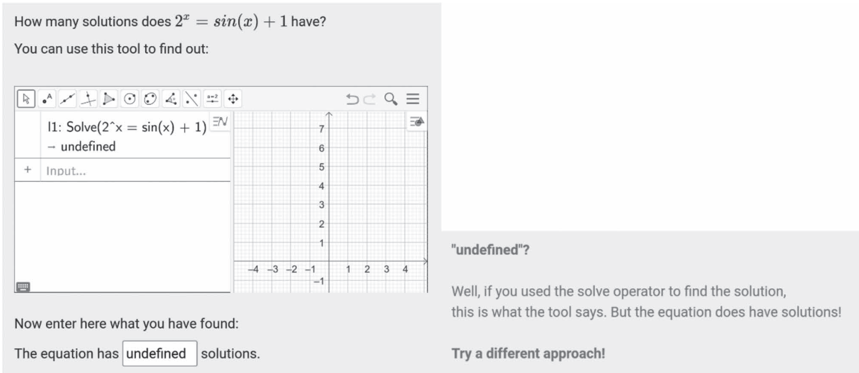


FIGURE 6.11 The algebraic 'solution'.

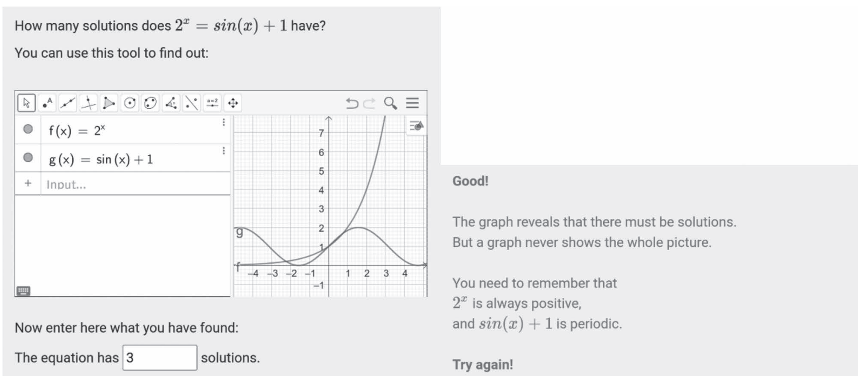


FIGURE 6.12 The graphical solution.

of the functions with f(x) = 2^x and g(x) = 1 + sin(x). Three solutions can be recognised as the x-coordinates of the intersection points of the graphs. On the one hand, this allows determining the solutions in the areas under consideration, to be determined numerically with an accuracy of a few decimal places. On the other hand, it must now be recognised that further solutions exist.

However, finding the other solutions to the equations then requires knowledge of the properties of the two functions. Even zooming in on a range of smaller negative x-values is a conscious process to be carried out by the learner (Figure 6.13). The correct answer to the question of the number of zeros – 'infinitely many' – requires mental thinking about the situation beyond the areas shown.

It is precisely for these insights that in-depth conceptual knowledge of the properties of the functions under consideration is required, but also of the typical mathematical way of thinking that goes 'to infinity'.

5.3.3 Quadratic equation with parameters: functional and object BMM

For which parameters b has the equation $x^2 + b \cdot x - 2 = 0$, $x, b \in \mathbb{R}$, one or two solutions?

The quadratic equation is first solved in the sense of the object BMM: It shows the general solution, depending on parameter b (Figure 6.14). However, the feedback now switches to the functional BMM by allowing to experiment with a slider and seeing the relationship between the graph of the quadratic function and the zeros of the function.

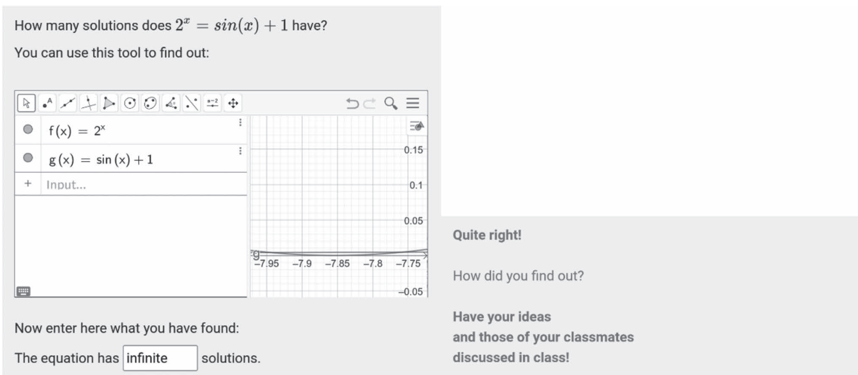


FIGURE 6.13 Looking for further solutions.

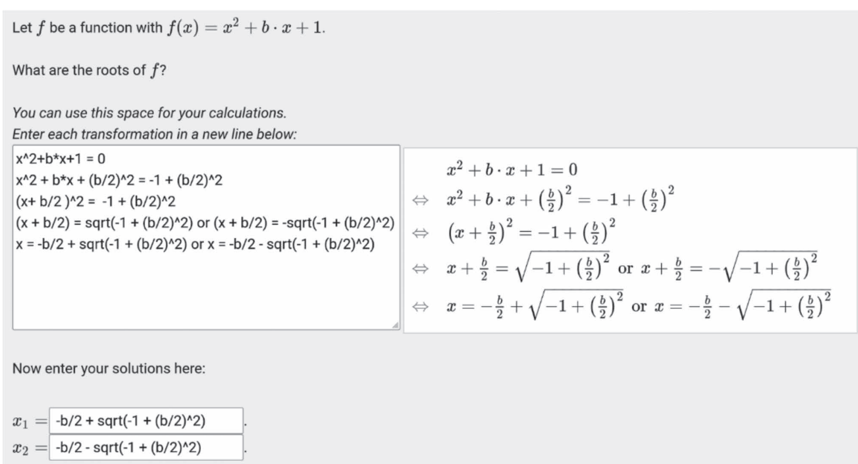


FIGURE 6.14 The AuthOMath file for the equation $f'(x) = ax(x-10)(x+10)$.

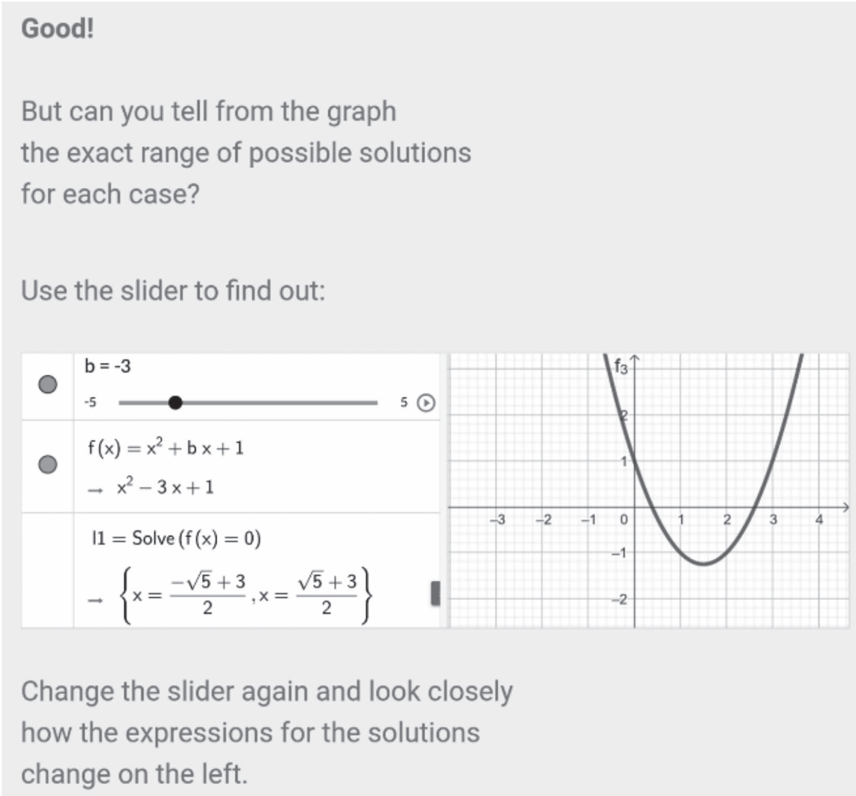


FIGURE 6.14 (Continued)

6 Conclusion

Digital assessments can support learning by providing the opportunity to present tasks and problems in a dynamic, interactive, and multi-presentational way. This allows learners to use different problem-solving strategies, for example, while using different representations to solve problems, to get hints about problem-solving strategies by experimenting in an interactive conversation with the system or considering special cases of a problem while dynamically changing the situation as appropriate. In addition, environmental or modelling problems can be authentically represented using, for example, films, simulations, or interactive applets.

While there is a whole range of proposals for the digital assessment of procedural knowledge, especially with computer algebra systems, such as Mathematica, Maple, GeoGebra, and STACK, the assessment of conceptual knowledge is far more complex. Both the questions and the expected answers need to be seen in the context of understanding mathematical concepts or algorithms. One possibility is to focus on basic mental models (BMMs) as the central core of conceptual understanding of mathematical concepts. However, the development of BMMs requires a long-term, demanding theoretical and empirical process, which must be followed by the development of adequate tasks.

In addition to the task and the task design, feedback – and the feedback design – is the second decisive aspect of formative digital assessment. For example, feedback can also explicitly ask about the conceptual background of the content being worked on in the context of tasks that relate to procedural knowledge. Adaptive feedback provides solution-oriented indications of the learner's solutions and supports an individualised learning process. Moreover, feedback can help overcome obstacles when solving tasks and provide hints on solution strategies.

Finally, the most pressing challenge of the formative assessment of CK is the development of a digital environment with an easy access for learners and practical application for teachers. On the technical level, STACK with integrated GeoGebra applets as provided by AuthOMath is a tool that promises to meet these requirements. The development of adequate content tasks and the empirical validation of these tasks will be a challenge for the coming years.

Reference list

Ball, L., Drijvers, P., Ladel, S., Siller, H.-S., Tabach, M., & Vale, C. (Eds.). (2018). *Uses of technology in primary and secondary mathematics education*. Springer. https://doi.org/10.1007/978-3-319-76575-4

Barana, A., Marchisio, M., & Sacchet, M. (2021). Interactive feedback for learning mathematics in a digital learning environment. *Education Sciences*, *11*(6), 279. https://doi.org/10.3390/educsci11060279

Barumbun, M., & Kharisma, D. (2022). Procedural knowledge or conceptual knowledge? Developing the so-called proceptual knowledge in mathematics learning. *Beta: Jurnal Tadris Matematika*, *15*(2). https://doi.org/10.20414/betajtm.v15i2.472

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74. http://dx.doi.org/10.1080/0969595980050102

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10/gcphxw

Bressoud, D., Martinez-Luaces, V., Ghedamsi, I., & Törner, G. (2017). Topic study group no. 16: Teaching and learning of calculus. In G. Kaiser (Ed.), *Proceedings of the 13th international congress on mathematical education* (pp. 447–452). Springer International Publishing. https://doi.org/10.1007/978-3-319-62597-3_43

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, *34*(4), 344–377. https://doi.org/10.1016/j.dr.2014.10.001

Drijvers, P., Ball, L. Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). *Uses of technology in lower secondary mathematics education. A concise topical survey*. Springer. https://doi.org/10.1007/978-3-319-33666-4

Drijvers, P., & Sinclair, N. (2023). The role of digital technologies in mathematics education: Purposes and perspectives. *ZDM Mathematics Education*, *56*(2), 239–248. https://doi.org/10.1007/s11858-023-01535-x

Duval, R. (1999). Representation, vision and visualization: Cognitive functions in mathematical thinking. Basic issues for learning. In F. Hitt & M. Santos (Eds.), *Proceedings of the twenty-first annual meeting of the North American chapter of the international group for the psychology of mathematics education* (Vol. I, pp. 3–26). Cu-ernavaca, Morelos, México.

Fahlgren, M., & Brunström, M. (2022). Example-generating tasks in a computer-aided assessment system: Redesign based on student responses. In H.-G. Weigand, A. Donevska-Todorova, E. Faggiano, P. Iannone, J. Medová, & M. Turgut (Eds.), *MEDA3 mathematics education in digital age 3. Proceedings of the 13th ERME topic conference (ETC13) held on 7–9 September 2022 in Nitra, Slovakia*. Constantine the Philosopher University in Nitra.

Fahlgren, M., Brunström, M., Dilling, F., Kristinsdóttir, B., Pinkernell, G., & Weigand, H.-G. (2021). Technology-rich assessment in mathematics education. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalová, & H.-G. Weigand (Eds.), *Mathematics education in the digital age* (pp. 69–83). Routledge. https://doi.org/10.4324/9781003137580-5

Gray, E. M., & Tall, D. O. (1994). Duality, ambiguity, and flexibility: A 'proceptual' view of simple arithmetic. *Journal for Research in Mathematics Education*, *25*(2), 116–140. https://doi.org/10.2307/749505

Greefrath, G., Oldenburg, R., Siller, H.-S., Ulm, V., & Weigand, H.-G. (2016). Aspects and "Grundvorstellungen" of the concept of derivative and integral. *Journal für Mathematikdidaktik*, *37*(S1), 99–129. https://doi.org/10.3102/003465430298487

Haapasalo, L., & Kadijevich, D. (2000). Two types of mathematical knowledge and their relation. *JMD – Journal für Mathematikdidaktik*, *21*(2), 139–157. https://doi.org/10.1007/BF03338914

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hechter, J., Stols, G., & Combrinck, C. (2022). The reciprocal relationship between conceptual and procedural knowledge – a case study of two calculus problems. *African Journal of Research in Mathematics, Science and Technology Education*, *26*(2), 111–124. https://doi.org/10.1080/18117295.2022.2101271

Hegedus, S. J., & Moreno-Armella, L. (2009). Introduction: The transformative nature of "dynamic" educational technology. *ZDM Mathematics Education*, *41*(4), 397–398. https://doi.org/10.1007/s11858-009-0201-9

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge* (pp. 1–27). Lawrence Erlbaum. https://doi.org/10.4324/9780203063538

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM Mathematics Education*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Hurrell, D. (2021). Conceptual knowledge OR procedural knowledge or conceptual knowledge AND procedural knowledge: Why the conjunction is important to teachers. *Australian Journal of Teacher Education*, *46*(2), 57–71. https://doi.org/10.14221/ajte.2021v46n2.4

Iannone, P., Lemmo, A., & Kinnear, G. (2025). The transfer of assessment from pen and paper to e-assessment. In E. Geraniou, C. Crisan, & M. Mavrikis (Eds.), *Digital technology and artificial intelligence in mathematics education assessment*. Routledge.

Klenowski, V. (2009). Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 263–268. https://doi.org/10.1080/09695940903319646

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Moreno, J., & Pineda, A. F. (2020). A framework for automated formative assessment in mathematics courses. *IEEE Access*, *8*, 30152–30159. https://doi.org/10.1109/ACCESS.2020.2973026

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning – not just for assessment. *Educational Psychology Review*, *35*(3), 89. https://doi.org/10.1007/s10648-023-09808-3

Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental Psychology*, *51*(3), 214–228. https://doi.org/10.1027/1618-3169.51.3.214

Pinkernell, G. (2024). Fachdidaktische Designparameter für ein automatisches Feedback auf Aufgabenebene. *Beiträge zum Mathematikunterricht*, *57*. https://doi.org/10.17877/DE290R-25016

Pinkernell, G., Diego-Mantecón, J. M., Lavicza, Z., & Sangwin, C. (2023). AuthOMath: Combining the strengths of STACK and GeoGebra for school and academic mathematics. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(3), 201–204. https://doi.org/10.3991/ijet.v18i03.36535

Ripley, M. (2009). Transformational computer-based testing. In J. Björnsson & F. Scheuermann (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 92–98). OPOCE.

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*, *77*, 20–26.

Skemp, R. R. (1986). *The psychology of learning mathematics*. Penguin.

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (Vol. 27, pp. 721–752). Springer. https://doi.org/10.1007/978-1-4614-4684-2_23

Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Vos, P. (2016). *Assessment in mathematics education: Large-scale assessment and classroom assessment*. Springer. https://doi.org/10.1007/978-3-319-32394-7

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*(2), 151–169. https://doi.org/10.1007/BF00305619

Vollrath, H.-J., (1984). *Methodik des Begriffslehrens*. Klett.

Vom Hofe, R., & Blum, W. (2016). "Basic conceptions" as a category of subject-matter didactics. *Journal Für Mathematik-Didaktik*, *37*(S1), 225–254. https://doi.org/10.1007/s13138-016-0107-3

Watson, A., & Mason, J. (2005). *Mathematics as a constructive activity: Learners generating examples*. Lawrence Erlbaum Associates.

Watson, A., & Ohtani, M. (Eds.). (2021). *Task design in mathematics education: An ICMI study 22*. Springer. https://doi.org/10.1007/978-3-319-09629-2

Weigand, H.-G., Greefrath, G., Oldenburg, R., Siller, S., & Ulm, V. (2017). Aspects and basic mental models ("Grundvorstellungen") of basic concepts of calculus.

In B. Kaur, W. K. Ho, T. L. Toh, & B. H. Choy (Eds.), *Proceedings of the 41st conference of the international group for the psychology of mathematics education* (Vol. 4, pp. 313–320). PME, Singapore.

Weigand, H.-G., Pinkernell, G., & Schüler-Meyer, A. (2022). Basic mental models of equations – theoretical conception and practical meaning. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the twelfth congress of CERME* (pp. 654–661). Free University of Bozen-Bolzano, Italy and ERME.

Yerushalmy, M. (2005). Functions of interactive visual representations in interactive mathematical textbooks. *International Journal of Computers for Mathematical Learning*, *10*(3), 217–249. https://doi.org/10.1007/s10758-005-0538-2

Yerushalmy, M., Olsher, S., Harel, R., & Chazan, D. (2023). Supporting inquiry learning: An intellectual mirror that describes what it "sees". *Digital Experiences in Mathematics Education*, *9*(2), 315–342. https://doi.org/10.1007/s40751-022-00120-3

Zuya, H. E. (2017). Prospective teachers' conceptual and procedural knowledge in mathematics: The case of algebra. *American Journal of Educational Research*, *5*(3), 310–315. https://doi.org/10.12691/education-5-3-12

7

ADVANCING AUTOMATED ASSESSMENT FOR OPEN-ENDED QUESTIONS IN MATHEMATICS

Sami Baral, Li Cheng, Anthony F. Botelho, and Neil T. Heffernan

1 Introduction

In mathematics education, questions are typically categorised as either close-ended or open-ended. *Close-ended questions* have a single or limited set of correct answers and a straightforward solution path, making these questions relatively easy to assess. In contrast, *open-ended questions* require students to generate their own solutions and explain their reasoning, often resulting in varied correct and incorrect responses, depending on the approach. Open-ended questions play a crucial role in fostering deeper understanding and critical thinking in mathematics (Kwon et al., 2006). They encourage students to articulate their reasoning, engage in mathematical discussions, and apply their knowledge to novel real-world situations (Lomibao et al., 2016). The emphasis on math communication fostered by these types of questions is vital for learning and applying mathematical concepts in real-world contexts, where questions are rarely presented with predefined solutions.

Historically, close-ended questions have dominated both traditional classrooms and online learning platforms. These questions are easier to assess automatically, as they allow for quick feedback and straightforward correction. Automated tools and educational technologies have thus predominantly supported close-ended assessments, providing students with immediate insights into their understanding of specific topics. However, the responses to open-ended questions are often diverse in content and can include different types of formats, including texts, diagrams, tables, mathematical formulas, and math expressions. The complexity and variation in these open responses present a unique challenge for assessment.

DOI: 10.4324/9781003533764-7

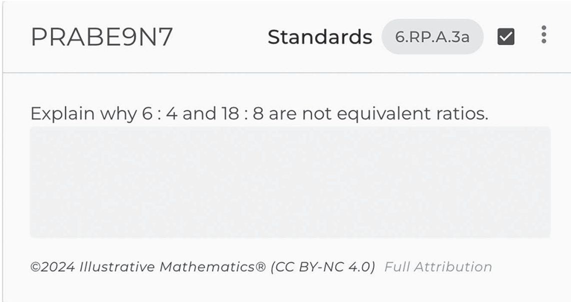


FIGURE 7.1 Example of an open-ended question taken from the Illustrative Mathematics curriculum within the ASSISTments platform. This is an open-ended question that asks students to explain why two given ratios are not equivalent, and this question can be correctly answered in varied ways.

The inherent complexity of open responses has limited the development of automated support systems for these types of questions. As a result, for open-ended questions, students have to wait for their teachers to review their responses and provide feedback, which is often delayed by days or weeks, or sometimes the students do not even receive feedback. The diverse and flexible nature of open responses, the presence of multiple correct answers, and the absence of a prescribed order of thought processes for open-ended questions demand significant time and effort from teachers and present significant challenges for traditional assessment.

Despite these challenges, the importance of open-ended questions is increasingly recognised, and contemporary mathematics curricula are integrating them more extensively. Popular open educational resource (OER) math curricula like Illustrative Mathematics feature approximately 40% of questions as open-ended. There has been a great need to develop automatic assessments for open-ended questions. With the development of AI-powered tools, automating the assessment of open responses becomes promising. To address the significant need to assess open responses, ASSISTments (Heffernan & Heffernan, 2014), an online learning platform for K–12 mathematics learning, has endeavoured to develop more inclusive automatic assessments and feedback methods for open-ended questions. In this book chapter, we delve into our research and the development of automated assessments in the ASSISTments platform for open-ended questions in mathematics. We further synthesise key insights and challenges in the process to shed light on future research and development of automated assessments in mathematics education.

TABLE 7.1 Examples of student responses and teacher scores and feedback to the open-ended question from Figure 7.1

Students' open responses	*Teacher score*	*Teacher feedback*
Because they don't add up correctly.	0	You don't add up ratios to see if they're equivalent.
They are not equivalent because 18:8 is bigger ratio.	0	Just because it is larger does not mean that it is not equivalent. Are they being multiplied by the same scale factor?
6:4 and 18:8 are not equivalent because 6 is equal to 18 when 18 is divided by 3, so the ratios are inequivalent.	2	Along the right lines, missing some steps. What about when 8 is divided by 4?
Because the greatest common factor for 6:4 is 3:2, while 18:8 is 9:2.	3	You are so close! The greatest common factor for 18:8 would actually be 9:4.
6:4 and 18:8 are not equivalent because to get from 4 to 8, you have to do 4 times 2; for 6 to get to 18, you have to do 6 times 3.	4	Great explanation.

Note: These examples are taken from the ASSISTments platform. Students' responses are paired with the scores and feedback from their teacher.

2 Advances in natural language processing and machine learning

Natural language processing (NLP) and machine learning (ML) are two significant branches of artificial intelligence (AI). NLP focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language. ML, on the other hand, provides algorithms and techniques that allow computers to learn from data and make predictions or decisions without being explicitly programmed.

NLP methods have evolved from simple, rule-based systems to sophisticated, deep learning models. Early methods in NLP relied on simplistic approaches, such as the bag-of-words approach, which counts occurrences of individual words within a document without capturing contextual or relational meanings. This approach, foundational in most of the early studies (Graesser et al., 2000; Sordoni et al., 2015), served as a baseline for comparison in evaluating student works. However, this approach could not interpret deeper semantic relationships or contextual nuances within the text. Techniques like n-grams were later introduced, which group sequences of words (bi-grams, tri-grams, etc.) to capture contextual information based on word proximity. However, these methods still fell short of capturing comprehensive relational understanding across text. Another pioneering method, term frequency–inverse document frequency (TF-IDF), weighted words based on their importance in a document relative to a corpus, enhancing the representation of meaningful keywords (Ramos, 2003).

The advances in deep learning introduced transformative approaches like Word2Vec and GloVe, which embed words into high-dimensional vector spaces to preserve semantic relationships. These models, developed by Mikolov et al. (2013) and Pennington et al. (2014), respectively, enable the encoding of nuanced semantic meanings and contextual understandings of language. These embeddings, learned from large-scale language corpora, can be applied across diverse contexts and enhance the robustness and applicability of NLP models in various tasks.

Moving beyond word-level embeddings, advancements in NLP extended to sentence- and document-level representations. Models like Doc2Vec (Le & Mikolov, 2014), Universal Sentence Encoder (Cer et al., 2018), and Sentence-BERT (Reimers & Gurevych, 2019) were developed to generate single-vector representations for entire sentences or documents. These representations capture broader semantic relationships and contextual nuances that span multiple words, clauses, and paragraphs, offering significant potential for improving the assessment of student open-ended questions.

3 Automated assessment models

The advances in natural language processing (NLP) and machine learning (ML) have led to the development of automated assessment methods in both mathematical and non-mathematical contexts. Methods such as C-rater (Leacock & Chodorow, 2003) use techniques to normalise student responses that vary across syntactic, morphological structure, pronouns, and synonyms to estimate the correctness of student responses to open-ended questions. Other researchers have explored clustering approaches to grade student responses (Basu et al., 2013, Brooks et al., 2014). These methods aim to identify common patterns and group responses into clusters, which can then be graded collectively. Clustering reduces the manual effort required for scoring while maintaining a level of consistency in assessment. However, these methods often require careful tuning and may struggle with outlier responses that do not fit neatly into predefined clusters. Deep learning models, particularly those based on neural networks, have shown remarkable success in capturing complex patterns in text data. Riordan et al. (2017) and Zhao et al. (2017) explored the use of high-dimensional representations of student work to assess the quality of responses. These models compare student responses to exemplar samples by leveraging the power of deep learning to handle the intricacies of language and context.

A study by Erickson et al. (2020) explored the potential of using NLP and machine learning methods to grade student responses to open-ended questions. Leveraging data from the ASSISTments platform, the researchers aimed to understand teacher scoring variations and develop models capable of accurately predicting teacher-assigned grades. Erickson et al. (2020) explored a range of machine learning approaches, including traditional methods, like random forest and XGBoost, as well as deep learning techniques, like long short-term memory (LSTM) networks. These models were trained on features extracted from student responses using NLP techniques like TF-IDF and word embeddings (bag-of-words and GloVe).

4 Semantic representation–based model

A significant advancement in NLP is the use of semantic representation models, which capture the meaning of sentences rather than just individual words. These models, including sentence transformers and embedding techniques such as SBERT (Reimers & Gurevych, 2019), allow us to evaluate the overall context and relevance of a student's response. Based on the semantic representation model, we developed a method called SBERT-Canberra for the automated assessment of open-ended questions in mathematics. We developed this method utilising a dataset of student responses to open-ended questions collected via the ASSISTments platform, along with the scores and feedback provided by teachers.

The SBERT-Canberra method relies on the contextual similarity of student responses. Each new student's answer is first compared to a set of historic answers in the dataset to find the closest match. This is achieved by using the Sentence-BERT model to generate a 768-dimensional vector representation for each student's answer. The similarity between student answers is then measured using the Canberra distance metric. The closest-matching answer is identified based on the smallest Canberra distance. The core idea behind this method is that by finding similar answers, we can utilise the assessment scores provided by teachers to previous similar student answers to score new student answers. Compared to previous works by Erickson et al. (2020), this method shows improved performance in assessing open-ended questions in mathematics and outperforms state-of-the-art methods.

However, recognising the limitations of this method and the challenges posed by variations in student responses, we conducted an exploratory error analysis to investigate areas where model predictions deviated significantly from teacher-provided scores. We extracted features across student responses and employed two regression models to observe absolute model error as the dependent variable.

With the first regression model, we explored features of student responses in the context of modelling error, identifying aspects that correlate most with

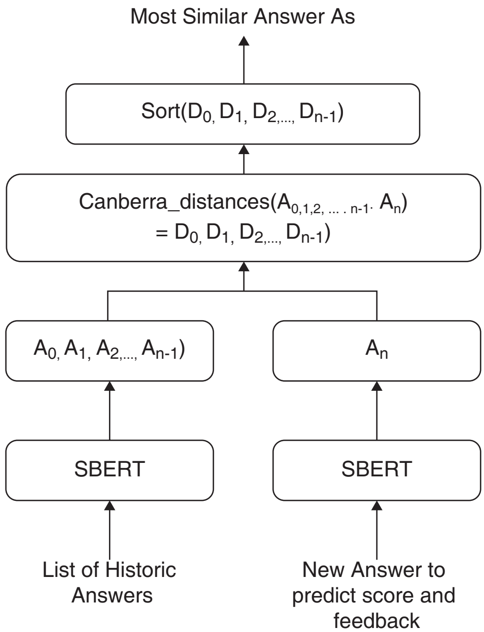


FIGURE 7.2 The SBERT-Canberra model for the automated scoring of student responses to open-ended questions in mathematics. This method is based on the contextual similarity of student responses and utilises scores provided by teachers to similar student responses.

Source: As proposed by Baral et al. (2021).

higher prediction errors. The results from the first regression model indicated that the presence of mathematical elements and images correlates with higher prediction errors for this model. The SBERT-Canberra method, based on the sentence-level representation of student responses, struggled to interpret complete mathematical terms and expressions. Additionally, since this model was designed for text-based responses, any images within the student responses were discarded.

The second regression model was a multilevel linear model that explored higher-level factors of student-, question-, and teacher-level identifiers that contributed to the model error. The results from this analysis suggested that question-level factors explained the majority of the variance of prediction error (beyond that of answer- and teacher-level factors). While the SBERT-Canberra method does consider individual questions, when producing its estimates by observing historical answers within each unique question, it appears that there are additional question-level factors not accounted for by this approach.

The findings from this analysis suggest the need for better representations of mathematics questions and methods for handling mathematical terms common in student open responses. The next section focuses on our model for mathematical terms.

5 Model for mathematical terms

The SBERT-Canberra model, which utilises pre-trained Sentence-BERT embeddings, outperformed earlier decision tree and deep learning approaches by leveraging large corpora of data to understand the semantic meaning of words and sentences. However, these embeddings often fail to recognise non-linguistic terms, like numbers and expressions, which are crucial in mathematical contexts.

To overcome this limitation, in our later work, the 'math term frequency' (MTF) method was proposed (Baral et al., 2022). The MTF method involves identifying the most frequent terms in students' answers and using these as features in a multinomial logistic regression model. This method parses student answers to identify non-linguistic terms and creates a frequency-based rubric for different scores. By ensembling the MTF method with the SBERT-Canberra model, we termed the model 'SBERT-MTF'. This model combines semantic representation with non-linguistic term matching. The SBERT-MTF model uses logistic regression to ensemble the score predictions from both the SBERT-Canberra and the MTF methods, which leverages the strengths of both models to improve the accuracy of automated assessments of open-ended questions. With the SBERT-MTF model, we were able to see improvements in the models' performance when compared to the SBERT-Canberra method. Through an observed error analysis, we were also able to demonstrate that the proposed SBERT-MTF model improves specifically in the presence of mathematical terms and expressions.

Similarly, Shen et al. (2021), developed the MathBERT model to address the limitations of traditional language models in handling mathematical terms and expressions. Building upon the base BERT (Devlin et al., 2018) architecture, MathBERT was trained on a large corpus of mathematical texts to understand mathematical language and terms. The pre-training for the MathBERT model involved datasets comprising mathematical texts, including textbooks, research papers, and educational materials from pre-K to high school to graduate-level mathematical curriculum, ensuring comprehensive coverage of the mathematical language. MathBERT demonstrated significant improvements in understanding and processing mathematical terms compared to general language models, resulting in superior performance on tasks requiring the interpretation of mathematical expressions and equations. In the context of automated assessment, MathBERT outperformed traditional models by accurately scoring student responses that included complex mathematical terms and expressions.

Zhang et al. (2022) further explored the use of in-context learning to enhance MathBERT's capabilities. This approach leverages the contextual information available in student responses to improve the model's performance. The in-context learning approach involves encoding both the textual and the mathematical context of student responses using MathBERT, capturing the surrounding text and any embedded mathematical expressions to provide a comprehensive representation of the response. For a given student response, MathBERT generates a context-aware embedding that is used to predict the score, capturing the interplay between the text and the mathematical content. The model showed a significant reduction in prediction errors, especially for previously unseen questions, demonstrating its generalisation capabilities across various mathematical tasks and datasets and highlighting its potential for broader application in automated assessment systems.

In addition to the model for mathematical terms, incorporating image recognition models is a valuable avenue to explore in building support for open-ended questions. In the next section, we discuss our models for analysing images in open responses.

6 Assessing visual responses

Visual responses are a common form of response to mathematical open-ended questions. Online learning platforms like ASSISTments allow students to upload images of their work on paper as part of their solution to open-ended questions. As suggested by our prior works, students often included images of their work with diverse formats, such as mathematical equations, graphs, and other visual representations, either alone or with typed text, when responding to open-ended answers. The presence of these image-inclusive responses significantly increased model error, as the models lacked a representation of visual responses.

In our more recent work (Baral et al., 2023), we employed deep learning–based image and text embedding methods to develop inclusive automated assessment tools that support both text- and image-based responses. Specifically, we utilised the CLIP (contrastive language–image pre-training) model developed by Radford et al. (2021), which is built on transformer architecture. This model encodes both natural language (text) and images into the same vector space through a multimodal pre-training approach. This approach is similar to the previously proposed 'SBERT-Canberra' method for text-based answers, except that this method uses a different embedding approach and can now support image-based answers. Using the CLIP model, we first encode textual and image responses into vector representations. For student responses containing both text and images, only the images are encoded in this method. After encoding all responses in the training data, we calculate the corresponding encoding for a new student answer (whether image- or text-based) and compare it to the embeddings in the training data.

The most similar response is then selected based on the shortest Canberra distance between the new response and those in the training set. The score is then predicted based on the score of the most similar response in the training set.

This preliminary analysis explored the feasibility of representing image-based student responses to address the limitations of existing methods. While this approach did not outperform the prior text-based SBERT-Canberra method, it demonstrated comparable performance. Furthermore, an error analysis similar to the one conducted for the previous method suggested that using pre-existing text and image embedding techniques can enhance the performance of auto-scoring models when images are present.

7 Future directions

In this book chapter, we have summarised our research and key findings for the assessment of open-ended questions in mathematics. We introduced methods to automate the assessment and examined aspects of the proposed methods that could be further improved. Following our findings that the proposed methods performed poorly in the presence of mathematical content and image-based student responses, we proposed solutions that demonstrated improved model performance. While these methods showed improvements in handling mathematical terms and image-based responses, there is significant potential for further advancement.

Future research should focus on enhancing the robustness and accuracy of automated assessment models by leveraging state-of-the-art advancements in natural language processing (NLP) and machine learning (ML). The rise of large language models (LLMs) such as GPT-4 (Achiam et al., 2023), LLaMA-3 (Touvron et al., 2023), and their variants presents a promising direction for future research and advancement in automated assessments for open-ended questions. These models have shown exceptional capabilities in understanding and generating human-like text, including mathematical language. Fine-tuning these LLMs for educational contexts can help them understand the nuances of student responses and the specific language used in mathematics, leading to better assessment systems. Additionally, LLMs can be utilised to provide richer, more granular feedback to students, helping them improve their mathematical language and reasoning, as well as helping them address any misconceptions and promote learning.

The multimodal capabilities of large language models offer another promising direction for automated assessment systems. Vision Language Models (VLMs) such as GPT-4 Vision are capable of understanding both complex language and images. Our initial success with the CLIP model, which encodes both textual and visual information, can be expanded upon by exploring newer and more advanced vision models like GPT-4 Vision for interpreting and analysing image-based student responses.

While this work primarily focused on providing numeric assessment scores for open-ended questions, another potential future direction is the development of automated feedback systems to provide meaningful feedback to students. Botelho et al. (2023) expanded the SBERT-Canberra method to provide textual feedback to students utilising a dataset of teacher-written feedback to student responses. Since LLMs show exceptional capability in generating human-like text, future research should focus on applying LLMs to enhance feedback systems for such responses. Developing LLM-powered tools can provide immediate, personalised feedback to students on open-ended questions, identifying any errors or misconceptions they have and helping them learn more effectively. Future research should focus on designing and fine-tuning adaptive feedback systems that not only score responses but also identify specific areas where students struggle. By leveraging LLMs, these systems can generate targeted feedback that addresses individual misconceptions and guides students through progressive learning pathways. Further integrating these feedback and assessment systems with online learning platforms can help create a better learning environment by saving teachers time.

As advances in LLMs and AI-based systems rapidly progress, it becomes crucial to ensure fairness in these systems. Open-ended questions allow students to openly express their answers, which also means that these answers are at high risk of exposing any personally identifiable information (PII) of students. For example, a student's ethnicity may be exposed within their answer, and this risk is increased with image-based responses. Exposing these data points when developing or fine-tuning a model could mean that these models could learn inherent biases based on them. Proper measures must be taken to remove PII from these data, ensuring they are not exposed to the model during training or fine-tuning. Further, researchers should focus on developing methods to detect and mitigate biases in automated assessment systems, ensuring that the models perform equitably across diverse student populations and do not disadvantage any particular group. Ethical considerations should also be at the forefront, with transparent reporting of model limitations and continuous monitoring to prevent any unintended consequences when developing automated assessment and feedback systems.

In summary, our research presented methods for the automated assessment of open-ended questions in mathematics, but there is considerable room for future improvement. Leveraging the advancements in LLMs and VLMs can significantly enhance the robustness and accuracy of the assessment methods for open-ended questions in the future.

Reference list

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., & Zoph, B. (2023). Gpt-4 technical report. *arXiv preprint*. arXiv:2303.08774.

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). *Improving automated scoring of student open responses in mathematics*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED615565

Baral, S., Botelho, A. F., Santhanam, A., Gurung, A., Cheng, L., & Heffernan, N. (2023). *Auto-scoring student responses with images in mathematics*. International Educational Data Mining Society.

Baral, S., Seetharaman, K., Botelho, A. F., Wang, A., Heineman, G., & Heffernan, N. T. (2022). Enhancing auto-scoring of student open responses in the presence of mathematical terms and expressions. In *International conference on artificial intelligence in education* (pp. 685–690). Springer International Publishing.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402. Association for Computational Linguistics

Botelho, A. F., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, *39*(3), 823–840.

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: Using clusters to grade short answers at scale. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 89–98). Association for Computing Machinery.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint*. arXiv:1803.11175.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805.

Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020). The automated grading of student open responses in mathematics. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 615–624). Association for Computing Machinery.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G., & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*(2), 129–147.

Heffernan, N. T., & Heffernan, C. L. (2014). The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*, 470–497. https://doi.org/10.1007/s40593-014-0024-x

Kwon, O. N., Park, J. H., & Park, J. S. (2006). Cultivating divergent thinking in mathematics through an open-ended approach. *Asia Pacific Education Review*, *7*, 51–61.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196). PMLR.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*, 389–405.

Lomibao, L. S., Luna, C. A., & Namoco, R. A. (2016). The influence of mathematical communication on students' mathematics performance and anxiety. *American Journal of Educational Research*, *4*(5), 378–382.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. arXiv:1301.3781.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29–48). Citeseer.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*. arXiv:1908.10084.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. (2017). Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 159–168). Association for Computational Linguistics.

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., Graff, B., & Lee, D. (2021). Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint*. arXiv:2106.07340.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint*. arXiv:1506.06714.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint*. arXiv:2302.13971.

Zhang, M., Baral, S., Heffernan, N., & Lan, A. (2022). Automatic short math answer grading via in-context meta-learning. *arXiv preprint*. arXiv:2205.15219.

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 189–192). Association for Computing Machinery.

8

A RECONCEPTUALISATION OF THE INSTRUMENTAL GENESIS PROCESS IN THE CONTEXT OF HUMAN–LLM INTERACTIONS

Exploring students' perspective on their use of GPT-4o as a tool to support self-assessment

Annalisa Cusi and Francesco Contel

1 Introduction

In this chapter, we address the challenging question of whether AI-based conversational agents can be considered as valid educational resources for promoting formative assessment (FA), with a focus on students' self-assessment practices in mathematics. Our study is framed within a Vygotskian perspective (Vygotsky, 1930–1934/1978), in which interaction with peers and experts plays a crucial role in students' learning, and is based on the hypothesis that FA and metacognition are deeply intertwined (Cusi et al., 2017). Given the linguistic dimension of metacognitive processes, real-time and human-like text generation, typical of conversational agents, could be an essential feature for the development of an appropriate tutoring environment.

Based on this assumption, we are investigating the potential role that a large language model (LLM) could play (or not) as a metacognitive tutor, that is, an expert who supports students in making their thinking visible and in promoting their metacognitive reflections on the strategies activated during problem-solving activities and on the learning process itself. The results of the first study (Contel & Cusi, 2024) suggested to us that this research should be developed together with the study of the interaction between a human user and an LLM as a consequence of the unique features of agency and interactivity of LLMs. There is little research on this topic. Yoon et al. (2024), for example, recently proposed a framework for describing human–AI interactions when dealing with mathematical proofs at the university level, emphasising the role of students' conceptions of both proof and AI as key factors in determining the effectiveness of the resource.

DOI: 10.4324/9781003533764-8

In this chapter we combine the study of both the user and the LLM perspectives by considering the instrumental relationship between users and artefacts. We discuss the significance of the main elements of the *instrumental genesis* process (Rabardel, 2002) in this context and propose a reconceptualisation of this process based on the identification of four main elements that need to be taken into account in order to characterise the peculiarities of this process in the context under study. At the heart of this reconceptualisation is the interplay between the non-deterministic behaviour of LLMs and the user's subjective adaptation to this indeterminacy. To represent this reconceptualisation, we introduce the metaphor of Brownian motion, a thermodynamic problem that is a paradigmatic example of stochastic dynamics resulting from the combination of *noise* and *drift* force. The instrumental genesis is therefore read as a process of *noise reduction* aimed at giving the motion an overall regularity.

The reconceptualisation is presented starting from the results of a study focused on analysing the perspectives of a group of students on their use of GPT-4o[1] as a tool to develop self-assessment practices in problem-solving activities.

2 Formative assessment in mathematics through digital resources

Our study is situated in the broad field of research on FA with digital resources (DR), where DR are conceived as tools for carrying out FA practices, understood as practices through which evidence of student performance is elicited, interpreted, and used by the three main agents of FA (teachers, learners, or their peers) to make decisions about next steps in instruction (Black & Wiliam, 2009).

As a theoretical lens for analysing FA practices with DR, we refer to the model outlined by Cusi et al. (2024) – the WHW model ('where-when-how' model).

The WHW model consists of three main elements:

1. The main areas where FA practices can be developed (the WHERE of FA)
2. The different moments in which teachers engage in FA practices (the WHEN of FA)
3. The main functionalities of digital resources that support FA processes, that is, communicating, analysing, and adapting (the HOW of FA)

As we are considering students' perspectives on the ways in which GPT-4o promotes self-assessment, we will elaborate on (1) and (3).

With regard to (1), the practices we are investigating are situated within two of the four main specific areas of FA practice, namely, FA area 3 (*fostering the quality of feedback*) and FA area 4 (*involving students in peer and self-assessment*).

Effective practices in areas 3 and 4 should promote students' reflections on the three central processes on which FA focus (Wiliam & Thompson, 2007): (1) establishing where learners are in their learning, (2) establishing where learners are going, and (3) establishing how to get there.

As regards the different contents of feedback, Hattie and Timperley (2007) propose a categorisation into four main levels: feedback *about the task* (which concerns how well a task is being performed), feedback *about the processing of the task* (which concerns the processes underlying the task), feedback *about self-regulation* (which refers to the way in which students monitor, direct, and regulate actions towards the learning goal), and feedback *about the self as a person* (which consists of positive or negative evaluations of the student).

Regarding (3), the communicating and adapting functionalities certainly seem to be prerogatives of the DR under study. In relation to the communicating functionality, the focus here is on *communication with technology*, according to the categorisation of Ball and Barzel (2018), that is, a communication that involves an interplay between the user and the technology. With regard to the adapting functionality, Cusi et al. (2024) refer to the three categories introduced by Stangl (2022): (1) *passive adaptivity of technology* (technology offers tasks that the user has to choose), (2) *active adaptivity of technology* (technology offers tasks and decides about the further learning path), and (3) *intelligent adaptivity of technology* (learning material is designed on the basis of a learner profile that is generated by the technology and constantly expanded).

The WHW model provides a general framework for dealing with the use of digital resources for FA purposes, but its application to describe the use of LLMs for FA highlights the need for further reflection, particularly at the HOW level. Indeed, in the case of LLMs, the communicating functionality could introduce new, unprecedented challenges, as we will highlight in the example discussed in the following sections.

Furthermore, the adaptive functionality needs to be discussed in terms of its effective affordances for the user: Although LLMs have a certain degree of adaptation to the user's requirements, it is unclear if and how the user's needs are taken into account.

To reflect on these emerging issues, in the following section we present an overview of what is already known about the use of LLMs for FA purposes.

3 The use of LLMs for FA purposes: state of the art

AI techniques and FA practices unsurprisingly share a defining characteristic: They rely heavily on the quality and quantity of accessible data to inform the effectiveness of their interventions. However, a critical feature of the actual development of FA practices is the fact that they are often, and more crucially, perceived as time-consuming and difficult to implement in practice. For this reason, many researchers have highlighted the potential that AI tools could have in this area (e.g. Hopfenbeck et al., 2023). Due to the supposed intelligent adaptivity of LLMs, the interaction between a chatbot and a student does not rely on resources designed by a teacher, so LLMs do not seem to require demanding practice for educators.

A review of the research literature on the use of AI for assessment purposes reveals that the primary objective of research studies in this area has been to reduce teacher workload, whether explicitly stated or not: The use of AI tools has so far been limited to mostly procedural tasks, such as automated scoring of summative tests or collecting data on student performance to inform teachers' next steps. Automated scoring of essays is a well-established natural language processing task, and many studies are moving in this direction, showing a preponderance of quantitative approaches to AI-based assessment (González-Calatayud et al., 2021). González-Calatayud et al. (2021) point out that most studies lack a proper theoretical framework, so it is not always clear to what extent the proposed methods should be considered as FA practices. The focus is often on technical issues of validity rather than placing the proposed methods in a clear pedagogical perspective, and this is also the case in studies concerning LLMs (Latif & Zhai, 2024).

However, FA requires much more: The interpretation of students' reasoning in open-ended tasks is essential to provide effective feedback that could trigger self-assessment processes in students. This feature links the problem of using LLMs for FA with the broader (open) problem of understanding mathematical statements using AI techniques.

In this respect, a number of relevant issues emerge from the literature review. Challenges are posed by the structural limitations of this technology, namely, the lack of understanding at the semantic level (Frieder et al., 2024) and the inevitable production of non-factual or biased information (Navigli et al., 2023). A relevant issue addressed by many studies is the comparison between human and machine understanding. Evidence has been provided for the current mismatch between LLMs and the human mind in the process of conceptualisation in mathematics (Suresh et al., 2023; Hankeln, 2024).

Although some features of these models will certainly be improved by pre-training on text corpora that are more representative of mathematical and educational situations, we share the concerns of Li et al. (2023), who doubt the possibility of LLMs to identify students' meaning-making practices. However, we believe that more research is needed to effectively address these concerns, both at a practical and a theoretical level.

In order to effectively use LLMs for mathematics tutoring as a FA practice, it is crucial to address the problem from multiple perspectives. On the one hand, the limitations of the current approach suggest that training corpora need to include data with real classroom interactions (Hankeln, 2024) and the need to improve the mathematical skills of LLMs (Dilling, 2024; Schorcht et al., 2024). On the other hand, a deeper understanding of how the fundamental actors within the FA processes (i.e. students and teachers) perceive the support process mediated by LLMs is needed in order to improve the quality of feedback from LLMs. It is on this last point that the reflections developed in this chapter focus. For this reason, the next section is devoted to introducing the instrumental genesis framework and to start a reflection on the role it might play in interpreting the interaction between humans and LLMs.

4 The process of instrumental genesis and its use for interpreting human–LLM interaction

The notion of *instrumental genesis*, introduced by Rabardel (2002), is based on two dialectics: on the one hand, artefact vs. instrument and, on the other, instrumentation vs. instrumentalisation. As noted by Trouche (2020), the use of this framework for analysing the impact of the integration of ICT in mathematics learning has led to the development of the *instrumental approach to mathematics didactics*, which emphasises the role of artefacts as essential components of learning processes, focuses on the analysis of instrumentation and learning as long-term processes, and proposes to consider an instrument as a living entity throughout the student's activity when facing mathematical problems.

Trouche (2020), drawing on Rabardel's (2002) remarks on the role of the dialectic between instrumentation and instrumentalisation, proposes this reformulation of the definition of *instrumental genesis*:

> [I]nstrumentation and instrumentalisation are two intrinsically intertwined processes constituting each instrumental genesis, leading a subject to develop, from a given artefact, an instrument for performing a particular task; the instrumentation process is the tracer of the artefact on the subject's activity, while the instrumentalisation process is the tracer of the subjects' activity on the artefact.
>
> *(p. 409)*

The notion of instrumental genesis has been extended by Hegedus and Moreno-Armella (2010) to take into account the context of dynamic technological environments. This extension includes

> the simultaneous co-actions between a user's use of a tool and a software environment's use of a tool, the feedback and reaction of a user being a certain process of utilisation, and internalisation of how the tool is manipulated, used by the environment, and then re-used by the user.
>
> *(p. 31)*

Hegedus and Moreno-Armella (2010) have also extended the two notions of instrumentation and instrumentalisation in light of this new conceptualisation of instrumental genesis:

> For instrumentation, we additionally define it as how co-actions with a tool shape the user's actions and understanding of the use of such a tool within, and with respect to, an environment. Instrumentalisation is extended to how the tool is shaped by the user (user's knowledge) and the environment, i.e., when the tool is manipulated by environmental factors following a user-input.
>
> *(p. 30)*

We would like to propose a comparison between the conditions that led to Moreno-Armella's extension and the current state of research. The instrumental approach was developed to cope with the introduction of technologies such as computer algebra systems in the classroom. It was extended because the dynamic nature of geometric objects and the relative organisation of commands in tools like GeoGebra made it necessary to specify the role of the environment in the instrumental genesis process, as a result of the increased agency of the resource.

The user and the digital environment with which he/she interacts could be interpreted as alternately playing the roles of actors and re-actors in performing actions. According to this interpretation, the utilisation schemes developed by students in their interaction with an LLM influence the way in which the LLM reacts, and at the same time, the development of these utilisation schemes is influenced by the way in which the LLM supports students' self-assessment.

The constantly evolving and changing nature of utilisation schemes, the users' lack of familiarity with specific LLMs, and the very rapid pace at which LLMs develop are some of the components that characterise the complexity of the instrumental genesis process in the context of human–LLM interaction.

For this reason, we believe that LLMs, with their non-deterministic features, bring as much novelty as DGEs did in the early 2000s. This idea was confirmed by the results of our first study (Contel & Cusi, 2024): Starting from the open question of whether ChatGPT could be able to promote metacognitive processes during problem-solving, we found some evidence of initial utilisation schemes that were strongly associated with significant differences in the approach students took in using ChatGPT and in their perceived effectiveness of the tool; however, we also noticed that a certain level of awareness in using ChatGPT was not directly associated with obtaining the desired support. In other words, ChatGPT introduces an element of randomness into the process that is not automatically dealt with by users, even when this issue is explicitly raised.

Thus, we felt the need to reconceptualise the process of instrumental genesis in light of the increased complexity that the advent of LLMs has brought to the study of human–DR interaction. To introduce our reconceptualisation, in the next section we present the results of a study developed with the aim of providing a snapshot of the long-term process of instrumental genesis in the context of one-to-one interactions between some students and GPT-4o, who plays the role of a mathematics tutor for FA purposes. By analysing this snapshot, we highlight the main elements on which our reconceptualisation is based, as they may help characterise the complexity of the instrumental genesis process in the context of human–LLM interaction.

5 A snapshot of the process of instrumental genesis in the context of human–LLM interaction: a study on the use of GPT-4o to support students' self-assessment processes

Our study focused on the use of GPT-4o as a tool to support students' self-assessment in mathematics, involving a group of 19 upper secondary school students (grades 9 and 10), who voluntarily participated in the study. The students come from different classes with the same type of curriculum (classical lyceum). Our aim was to investigate the influence of the characteristics of human–LLM interactions on the processes of instrumentation and instrumentalisation by focusing on the following research questions (R1 on the instrumentation process and R2 on the instrumentalisation process):

RQ1. How do students interpret the role of GPT-4o in providing them feedback and support for self-assessment?
RQ2. How do students interpret their use of GPT-4o as a tool to receive feedback and support for self-assessment?

Most of the students had little or no experience of using GPT-4o at the time of the study, with the exception of two who described themselves as 'used to using it', but not in mathematics.

They were asked to work individually on two problem-solving tasks involving the exploration of numerical situations, the development of conjectures, and the construction of proofs.

Students were asked to use GPT-4o (the latest version available at the time of the study) as a mathematics tutor for the first task and to solve the second task with the help of an expert human tutor.

GPT-4o was prompted with a set of customised instructions informing the system of the context and specific requirements, namely, the age of the participants, the desired tutoring, and the feedback style (the prompts are available in the appendix).

	Before T1	After T1 (GPT-40 tutoring)	After T2 (Human tutoring)
Where is the student?	Have you ever done this kind of activity before? Do you think you have the skills to do it?	Did your interaction with GPT-40 help you understand whether the answers you gave were valid or not? If so, why?	Have you changed your mind about your ability to solve this kind of problems? If so, why?
Where should the student go?	Do you think this activity requires you to use algebraic skills? In what way?	Did GPT-40 help you to clarify the aim of the task?	After solving the second task with our help, would you answer the previous questions differently? If so, why?
What should the student do to get there?	What can you do to improve your algebra skills?	After doing the activity with GPT-4o, what do you think is the aim of this activity?	Having solved the second task with our support, how do you think you could improve your skills in this area? Why?

FIGURE 8.1 An excerpt from the semi-structured interview.

Students were asked to think aloud during their interactions with GPT-4o, with the researchers acting as observers, intervening only at moments of impasse or to elicit further explanation of thought processes. Semi-structured interviews were conducted with each student before and after their work on the two tasks. Each session lasted approximately 100 minutes.

The data collected consisted of audio and video recordings of the students working on the two tasks and of the semi-structured interviews.

In designing the semi-structured interviews, we explicitly referred to our theoretical framework on the use of DR for FA. In particular, we framed perceived feedback in terms of the three central processes on which FA should focus, by considering the following basic questions 'Where is the student?', 'Where should the student go?', 'What should the student do to get there?'. Examples of questions asked at different stages of the semi-structured interviews are shown in Figure 8.1.

6 Data analysis and results of the study

In the following, we will focus on the analysis of the collected data that we used to address the two research questions. As we are investigating students' perceptions of the GPT-4o feedback and its support of self-assessment at the same time as, or shortly after, the phenomena under investigation, we chose to conduct the analysis according to the interpretative phenomenological analysis (IPA) approach (Smith & Osborn, 2003), as it aims to analyse in detail how people involved in certain phenomena perceive and make sense of them.

As Smith and Osborn (2003) point out, the IPA approach aims to understand the complexity of meanings rather than to measure their frequency. This complexity can only be understood through a sustained engagement with the available data and a process of interpretation. For this reason, the best way to collect data for an IPA study is to use semi-structured interviews, as this allows the researcher to engage in a dialogue with the participants, modifying the questions in response to the participants' answers with the aim of exploring interesting ideas that emerge during the interviews.

The data analysis was developed in three main phases:

- *First phase*. The two authors individually read four interview transcripts (two for each researcher), annotating interesting ideas related to students' use of GPT-4o as a tool for feedback and self-assessment and emerging themes related to these ideas. In this phase, potentially interesting excerpts for the study were identified.
- *Second phase*. A systematic discussion was conducted between the two authors to share the first different emerging themes and the connections between them. At the end of this phase, different clusters of themes were identified; in the following, we will present the three clusters of themes related to our research questions.
- *Third phase*. The transcripts of the other interviews were subjected to further analysis in order to identify additional evidence relating to the clusters of themes that had emerged. This phase of analysis was carried out with reference to the findings of the previous phases. We sought to identify instances of alignment and misalignment in the students' perspectives in order to cover as wide a range of aspects as possible in relation to the research questions.

We present three main clusters, within which we have grouped the themes that allow us to answer the two research questions by characterising students' interpretation of the role of GPT-4o in providing feedback and support for self-assessment (RQ1) and their use of GPT-4o for the same purposes (RQ2).

The first cluster of themes relates to students' perceptions of the characteristics of GPT-4o feedback and the ways in which it supports self-assessment. This theme, therefore, relates to the first stages of students' exploration of GPT-4o within the process of *instrumentation*.

Most of the students describe GPT-4o's feedback as timely ('[A]rtificial intelligence . . . often responds directly and immediately to what it is asked') and, above all, responsive to the difficulties manifested by them. With regard to the responsiveness of GPT-4o's feedback, many students show that they particularly appreciate its frequent feedback about the task, manifesting a vision of feedback as a step-by-step explanation of what is right and what is wrong ('[W]hen I arrived at a first result, it would tell me whether it was right or not and then suggest how to continue'). The students who say they most appreciated the feedback about the processing of the task criticised the over-structured support provided by GPT-4o as it didn't allow them to work autonomously and reflect on their difficulties ('[W]hen I use GPT-4o it tells me in a more direct way what steps to take. . . . [A]t the moment it helps me more but then I think less, so in the long run it's more useful to have a little hint so I can think').

Some students also noted that GPT-4o sorts out information and does not give feedback on all the text written in the chat. In some cases, they interpret this fact as criticism; in other cases, they justify this behaviour by interpreting the lack of reaction from GPT-4o as implicit feedback, hypothesising that what it does not comment on is always right.

Regarding the support provided by GPT-4o for self-assessment, most of the students are more critical, stressing that GPT-4o is not clear about the objective of the path it has the student take, does not understand the student's difficulties, and is therefore less effective than a human tutor in supporting those students who have difficulties and, in particular, in providing long-term support ('[I]t doesn't give you a method, because it gives you very immediate answers, . . . answers for their own sake, that is, it gives you answers related to the question you ask, strictly related to that occasion').

Some students attribute the reasons for GPT-4o's ineffectiveness in providing support in case of difficulties to the lack of information in GPT-4o's possession, since it only receives the information that the students give it. In general, most students explain that GPT-4o is much less effective than a human tutor in supporting self-assessment, which therefore has to be developed autonomously by the students:

> I made this mistake and it seems to me . . . I realised it by myself . . . and then I told GPT-4o and then there . . . the situation changed a bit, but I realised the mistake by myself. Maybe, if there had been a human tutor, he would have told me, maybe without directly telling me the error. . . . [S]urely there would have been a help compared to GPT-4o, which there can't be, it's almost impossible.

When reflecting on this issue, many students note that one of the reasons for this lack of support in self-assessment is the ineffectiveness of GPT-4o in promoting metacognitive processes, especially in the case of students who have difficulties in mathematics:

> If I don't understand something, or if someone else doesn't understand something, I find it difficult to understand how else to ask GPT-4o how to explain it, that is, if I don't understand something, I also have to find a way to ask you how to explain it to me in a different way.

A final theme belonging to this cluster is that of GPT-4o's hallucinations and the loops into which it can lead students when solving mathematical problems, an aspect that makes GPT-4o a sometimes-unreliable tool. This theme emerges in particular from the interviews with students who claim to have experience of using GPT-4o. These students give a few examples that testify to the indeterminacy of the action–reaction processes that characterise their interaction with GPT-4o, stressing that its response cannot be predicted, regardless of how the question is formulated:

> [S]ometimes it happens that I'm specific and it concentrates on that, or I ask a question and it ranges, and other times when I ask it something, instead it starts to range in an incredible way, you don't know how . . .

Most of these students explain that they do not trust GPT-4o as a tool to control computations, and that the user has to decide to change the approach to a problem when he/she realises that the interaction is not productive due to GPT-4o's loops:

> You have to be careful with the calculation part because it can correct you even if it then says that it is actually wrong in its corrections . . .

> I changed my way because I could see that I was not coming to a conclusion. . . . even when I asked it 'sorry, I didn't understand, can you tell me why?', it always ended up, maybe by explaining a little more, always saying the same thing it had said before, so we were at a dead end.

The <u>second cluster</u> of themes relates to <u>students' perceptions of the characteristics of their ways of interacting with GPT-4o in order to promote its feedback and support for self-assessment</u>. This theme, therefore, refers to the initial development of students' utilisation schemes or their consolidation (in the case of students who are more experienced in using GPT-4o) within the process of *instrumentalisation*.

The students who say that they used GPT-4o for the first time during their interview emphasise first of all that they had to repeat the same question to GPT-4o in order to get useful feedback and that they realised the need to formulate questions in a straightforward way, since the way in which the questions to GPT-4o are formulated affects the type of answer one gets ('[W]hen what it had to answer me was a little less direct, it had more difficulties and also needed more time, more answers to get to what I needed').

Some of these students reflect on the communicative ambiguity that characterises the interaction with GPT-4o and on the language that should be used during this interaction, suggesting the need to use a language that can be unambiguously interpreted, such as mathematics. They contrast the approach of a human tutor, who understands beyond the mere written words or representations used, and the approach of GPT-4o, which interprets in a more 'literal' way what is written:

> [A] human tutor can understand better, even if there are some mistakes in the communication or something is not so clear, whereas GPT-4o, being a machine, let's say, only takes in very precise information.

Some students note that the effective use of GPT-4o as a self-assessment tool requires a high degree of autonomy on the part of the student, since it is necessary to be aware of one's own difficulties and to correctly identify one's own errors in order to ask questions and receive the right support:

> With a human tutor, it's easier for the person you're interacting with to understand your problem . . . whereas with GPT-4o, it's me who has to identify the problem first and then. . . . I mean, I have to be aware of what I did wrong in order to ask him certain questions.

A related theme is that of control: Some students observe that, when interacting with GPT-4o, control is exercised by the user because GPT-4o only responds when asked. According to them, this feature makes GPT-4o a less effective tool for self-assessment, since the user has to be in control of self-assessment processes in order to promote GPT-4o's activation of these processes: '[I]t's also a negative fact, the fact that I'm in control, because . . . if I, who's in control, don't understand the problem, how will GPT-4o, which is controlled by me, understand the problem?'

For this reason, the students who claim to have experience in using GPT-4o suggest that it should be used as a diagnostic tool or as a tool to carry out a given procedure, rather than as a tool to support the solution of complex problems. These students report a contrast between the stability[2] of the schemes they have developed to deal with simple tasks (such as asking for information or overcoming moments of impasse in communication) and a lack of stability of the schemes to select information and establish goals (Vergnaud, 2009) in the case of complex tasks (such as requiring support in interpreting symbolic expressions or in activating anticipating thoughts).

The <u>third cluster</u> is transversal to the others and concerns <u>students' a priori beliefs about the characteristics of GPT-4o in relation to its use as a self-assessment tool</u>. A first theme that emerges transversally from the students' interviews is the humanisation of GPT-4o due to the way it interacts with the user ('It almost feels like talking to a person because of the precision of the answers').

Another theme is the infallibility of GPT-4o, which, on the one hand, makes students take responsibility for the times when it did not really help them and, on the other hand, makes them feel reassured by the conviction that GPT-4o does not make mistakes and, therefore, when properly stimulated, can provide effective help: ‘I tend to trust it more, for a simple fact that I often see ChatGPT as a machine, so something also mathematical, something that is like precise, rational.’

Another theme in this cluster, which emerged in some of the interviews, relates to the advantages and disadvantages, from an emotional point of view, of interacting with GPT-4o compared to interacting with a human tutor. Indeed, some students highlighted GPT-4o’s lack of empathy on the one hand, contrasted with students’ lack of fear of being judged when interacting with it on the other. This perception of not being judged makes students feel free to ask more questions than when interacting with a human tutor.

7 Discussion: reconceptualisation of instrumental genesis in the context of human–LLMs interaction

The snapshot presented in the previous sections and the results of the analysis developed within our study allowed us to highlight four elements that characterise a possible reconceptualisation of the instrumental genesis process to take into account the complexity that the emergence of LLMs has brought to the study of human–DR interaction. In this section, we present our reconceptualisation in light of the results presented, by introducing the four elements. The first two elements introduce internal (first element) and external (second element) factors that influence instrumental genesis in human–LLM interaction. We use the terms *internal* and *external* in relation to the users’ perspective. The third and fourth elements refer to the characteristics of the human–LLM interaction. The third element introduces the need to reconceptualise the notion of co-action in light of the ways in which internal (human) and external (LLMs) factors influence this bidirectional process. The fourth element introduces a new way of characterising utilisation schemes, emphasising their unstable nature.

We then introduce the idea of *noise reduction* as a metaphor to represent our reconceptualisation of the instrumental genesis process.

7.1 First element: anthropomorphisation of the artefact and users’ metacognitive skills and experience with it as internal factors affecting instrumental genesis

In the previous section, within the third cluster of themes that emerged from our analysis, we introduced the tendency of users to *anthropomorphise* GPT-4o, attributing to it characteristics that are typical of human–human conversations, such as the widespread idea of ‘communication misunderstandings’, or the idea that GPT-4o lacks abilities such as empathy or a judgemental attitude. This phenomenon has been largely documented in cognitive psychology in the study of human–computers interaction (e.g. Nass et al., 1994; Epley et al., 2007), but to our knowledge it is not widely addressed in the research studies dealing with AI in mathematics education.

Another aspect, which emerged within the second cluster of themes, is the role of the users' metacognitive skills and knowledge of the artefact in strongly influencing their interaction with the LLM. Some students show, in their interview, that they are aware of the fact that GPT-4o provides a scaffolding that is only local and much too structured, lacking a long-term educational goal. Self-regulation processes are hardly stimulated by GPT-4o, as we can see from the feedback, which is mostly at the task level. In relation to this, students also note that only those who are already aware of their fragility can effectively question GPT-4o and receive feedback in line with the perceived difficulty; those who are not will blindly follow the chatbot's lead. This confirms the results of a previous study (Contel & Cusi, 2024) in which we associated a phenomenon not typical of social interactions, namely, the introduction of deliberate pauses in the flow of the chat to encourage self-questioning and reflection, with a higher level of utilisation schemes.

Indeed, it is impressive how the sustained support provided by GPT-4o is perceived as effective – meaning, that users' trust in the AI's mathematical capabilities is not challenged, despite the various evidences in the interactions and the critical perspectives suggested in the interviews. This phenomenon of associating mathematics with AI algorithms and vice versa is an attempt of sense-making of complex entities ('algorithmic imaginary', Bucher, 2016). It is somehow complementary to the aforementioned tendency towards anthropomorphising; these two tendencies could be read in terms of the wider lenses of users' beliefs.

7.2 Second element: uncertainty and randomness in the use of the artefact as external factors affecting instrumental genesis

LLMs are inherently probabilistic. Users are mostly unaware of this, but in their interview, they show an awareness of the need to find effective ways of questioning GPT-4o to get useful answers, that is, they reflect on their instrumentalisation process.

The interviews clearly show what could be seen as the flowering of 'theorems-in-action': questions need to be direct, specific, using formal language to reduce ambiguity in communication. Repeating questions and using specific strategies (e.g. copy–paste) are attempts to steer the chatbot in the desired direction.

Another important finding is that expert users are aware of the possibility of hallucinations or conversational loops leading to dead ends. This leads them to think about the 'right way' to use this resource; they explicitly consider that different tasks are of different complexity and that AI can only handle some of them or some aspects of them. However, experience may not be enough: The need for more control in the interaction is evident in some cases where users complain about the structural randomness of the tutor. The interviews also revealed the influence of certain beliefs about GPT-4o's mathematical competence (mentioned as internal factors affecting human–LLMs interaction), and consequently about the quality of the feedback provided, on the development of students' awareness of GPT-4o's hallucinations or conversational loops. Even factual evidence of misleading interventions does not challenge the idea that 'machines understand mathematics because they share the same logic'. Expert users express concerns about trust but are not immune to this belief.

7.3 Third element: indeterminacy of the co-actions between the user and the environment

Our results have highlighted the need to reconceptualise the notion of co-action between a user and an environment (Hegedus & Moreno-Armella, 2010) in the context of human–LLM interaction, since the GPT-4o environment, despite the users' efforts at predictability described in the discussion of the second cluster of themes, and despite the back-and-forth dialogical structure of co-action, does not allow the environment's response to be uniquely associated with the user's action, as some of the students who participated in our study pointed out. In fact, GPT-4o reacts, but not in a deterministic way, which prevents the user from developing a reproducible and functional set of actions. This aspect strongly characterises the instrumental genesis and has an impact on the specificity of the utilisation schemes that emerge from this process.

7.4 Fourth element: liquidity as a characteristic of the emerging utilisation schemes

A certain degree of adaptability is required as a feature of utilisation schemes when a user interacts with an artefact, because, as Rabardel (2002) points out, utilisation schemes are constantly evolving through 'adaptation, combination, coordination, inclusion and reciprocal assimilation, the assimilation of new artefacts into already constituted schemes' (p. 103).

In the context of human–LLM interaction, we ascribe the property of *liquidity*[3] to utilisation schemes, in line with the users' perception of the lack of stability of schemes in the case of complex tasks. In fact, on the one hand, certain tasks (asking for information, performing simple calculations, summarising, formatting) may correspond to relatively stable schemes because the support requested is more likely to be useful. On the other hand, more complex tasks (such as those tested in our study) can't be performed with a fixed approach that can be associated with a scheme that becomes stable over time. In this second case, the inability to develop stable schemes makes them liquid due to their uncertain nature.

The internal and external factors discussed previously exacerbate this aspect since they lead to a need for adaptation on the part of the user, which seems more drastic compared to the instrumentation processes that occur in the interaction with other types of artefacts. In the case of non-expert users, this adaptation becomes even more delicate.

7.5 A metaphorical sketch of the process: instrumental genesis as noise reduction

We introduce a mathematical metaphor that attempts to capture the essence of the complexity involved in instrumental genesis in the context of human–LMM interaction, in light of the four elements introduced in the previous sections. The need to characterise this complexity and the ways in which a user might try to dominate or at least control it led us to look to the approaches developed in physics for modelling complex systems.

In the field of stochastic processes, the first attempts to describe non-deterministic dynamics were made by Einstein, Smoluchovsky, and Langevin, among others, arising from the famous problem of Brownian motion. In modern ('physicists'') notation, we consider a process obeying Langevin's stochastic differential equation:

$$m\frac{dx}{dt} = -\gamma x + \eta(t)$$

where $-\gamma x$ is the (deterministic) force and $\eta(t)$ is a gaussian noise with variance σ^2, that is, a measure of the size of the fluctuations; the meaning of the equation can be interpreted as follows: The trajectory of the particle depends on two contributions, a deterministic one (the *drift*) and a random one (the *noise*).

It's easy to see the effect of the drift on the trajectory by looking at Figure 8.2, which shows a 2D system undergoing Langevin dynamics. The pink curve shows a realisation of this process without drift, while the orange curve has a 'strong' drift contribution. The curves refer to the same number of time steps. The drift introduces a general direction that is still affected by fluctuations, but the overall motion has some regularity, and the distance covered in time improves significantly as the drift contribution is stronger.

According to our interpretation of this metaphor, the instrumental genesis process is influenced by various factors that act as *noise*:

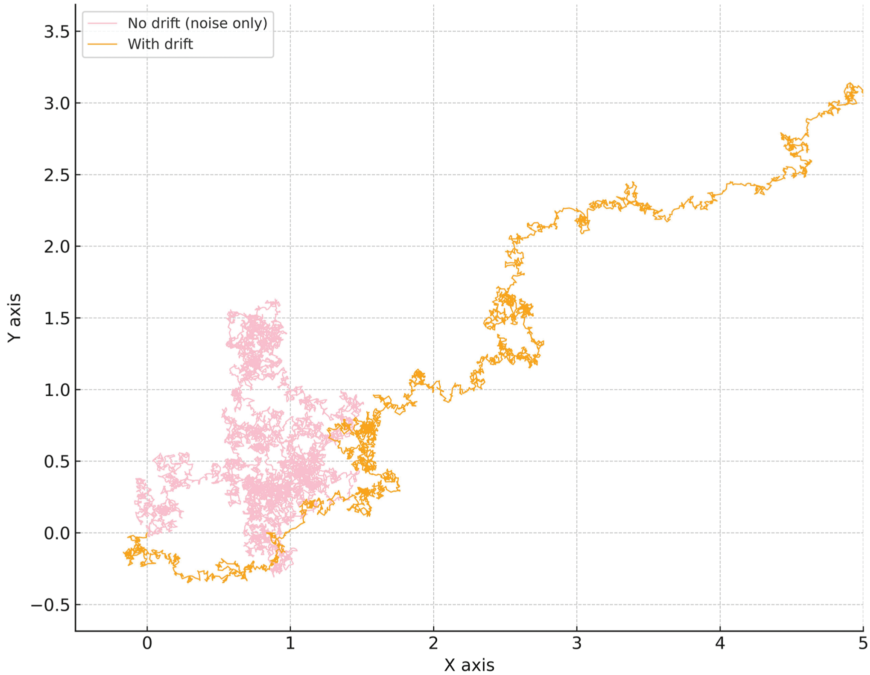


FIGURE 8.2 Comparison of 2D Langevin dynamics with different drift magnitudes.

- External factors, such as the uncertainty and randomness introduced into users' processes supported by LLM due to LLM's structural limitations (second element in our reconceptualisation)
- Internal factors, such as the anthropomorphisation of LLM tools and specific beliefs that lead the users to blindly trust LLM's capabilities, due to their perception of LLMs as autonomous social actors with whom they interact (Formosa, 2021), and to develop pseudo-theories about the 'logic behind the scene' (first element in our reconceptualisation)
- The characteristic of indeterminacy of the co-actions between the user and the LLM environment (third element in our reconceptualisation), which is a result of the interaction of the external and internal factors earlier described

This *noise* prevents users from establishing stable schemes (fourth element in our reconceptualisation), which is reflected in the erratic trajectory that only follows a certain direction on average. A *drift* is therefore needed to reduce the instability caused by this noise. The inability to control the external factors influencing the co-action process suggests the need to focus on internal factors as a way of helping users to direct to the instrumental genesis process towards more stable schemes.

7.6 Conclusion

In this chapter, starting from a study focused on students' perspectives on their use of GPT-4o for self-assessment in mathematics, we have reflected on the complexity that the use of LLMs brings to the analysis of the interaction between humans and digital resources, suggesting a possible reconceptualisation of the instrumental genesis process in this particular context.

As it makes it possible to highlight specific factors that strongly influence the dynamics of co-action and the development of utilisation schemes during human–LLM interaction, this reconceptualisation suggests different actions that specific stakeholders could activate in order to promote a more efficient use of LLMs as self-assessment tools and, more generally, as tools to support problem-solving processes in mathematics.

First of all, highlighting the key role played by uncertainty in the use of LLMs and by the indeterminacy of co-action mechanisms in negatively affecting the use of LLMs as educational self-assessment tools suggests that a first level of actions could be performed by designers and developers in trying to reduce the components of noise brought by these external factors. For this reason, since LLMs were not conceived for educational purposes, more discussion and collaboration between designers, developers, and educational researchers is desirable.

At the same time, our reconceptualisation allows us to identify specific internal factors that influence the co-action mechanisms and the development of the user's utilisation schemes, such as the anthropomorphisation of LLMs and the user's often unconditional trust in the LLM's capabilities. Reducing the noise introduced by these factors should be a major concern for the educational use of LLMs, especially if we focus on self-assessment. Indeed, the fact that GPT-4o is perceived as a social actor unbalances the communication: Does the student interacting with it really perceive that he/she is doing self-assessment? If a conversational agent is perceived as an always-effective subject to interact with, this may mean that the learner does not activate the introspective processes necessary for self-assessment. This is consistent with recent studies on autonomy in human–machine interaction (Formosa, 2021). Our study suggests that the noise introduced by these factors could be reduced if the users were supported in controlling their influence on the interaction with LLMs. In particular, the role played by the development of metacognitive acts seems to be relevant, as evidenced by some of the reflections presented within the first cluster of themes that emerged from the students' interviews. In this way, metacognitive skills could play the role of the *drift* term that steers the movement, that is, the instrumental genesis process, in the desired direction.

These reflections suggest possible actions that teachers could take to reduce these components of noise. On the one hand, in order to counter the consolidation of specific beliefs that lead to the anthropomorphisation of LLMs or to an unconditional trust in their capabilities, it is important that teachers support students' reflections in order to make them become more aware of the ways in which they conceive LLMs and interpret their interaction with them. On the other hand, in order to promote students' development of metacognitive skills useful for controlling and directing their interaction with LLMs, it is desirable that teachers promote systematic collective reflection on the ways in which LLMs could be used for self-assessment purposes and on the reasons behind the most effective strategies that could be implemented to achieve these goals.

Notes

1 GPT-4 is one version of the popular chatbot ChatGPT developed by OpenAI. For further info about the different models: https://chatgpt.com/.
2 *Stability* is to be interpreted as a property of the scheme–situation pair: We consider the scheme developed by a user to be *stable* if it leads to a foreseen and desired (but not necessarily always effective) outcome in a given situation.
3 *Liquidity* is to be interpreted, very broadly speaking, in the sense popularised by Zygmunt Bauman in many of his works.

Reference list

Ball, L., & Barzel, B. (2018). Communication when learning and teaching mathematics with technology. In L. Ball, P. Drijvers, S. Ladel, H.-S. Siller, M. Tabach, & C. Vale (Eds.), *Uses of technology in primary and secondary mathematics education: Tools, topics and trends* (pp. 227–244). Springer International Publishing. https://doi.org/10.1007/978-3-319-76575-4_12

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bucher, T. (2016). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Contel, F., & Cusi, A. (2024). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*. https://doi.org/10.1007/s40751-024-00164-7

Cusi, A., Aldon, G., Barzel, B., & Olsher, S. (2024). Rethinking teachers' formative assessment practices within technology-enhanced classrooms. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 1–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-95060-6_29-1

Cusi, A., Morselli, F., & Sabena, C. (2017). Promoting formative assessment in a connected classroom environment: Design and implementation of digital resources. *ZDM Mathematics Education*, *49*, 755–767. https://doi.org/10.1007/s11858-017-0878-0

Dilling, F. (2024). Large language models as formative assessment and feedback tools? – A systematic report. In P. Iannone, F. Moons, C. Drüke-Noe, E. Geraniou,

F. Morselli, K. Klingbeil, M. Veldhuis, & S. Olsher (Eds.). *Proceedings of FAME 1 – Feedback & assessment in mathematics education (ETC 14)*. Utrecht University and ERME. https://doi.org/10.5281/zenodo.14231455

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864. https://doi.org/10.1037/0033-295X.114.4.864

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines*, *31*(4), 595–616. https://doi.org/10.1007/s11023-021-09579-2

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, *36*. https://doi.org/10.48550/arXiv.2301.13867

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467. https://doi.org/10.3390/app11125467

Hankeln, C. (2024). Challenges in using ChatGPT for assessing conceptual understanding in mathematics education. *Journal of Mathematics Education*, *17*(1), 1–15. https://doi.org/10.26711/007577152790171

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hegedus, S. J., & Moreno-Armella, L. (2010). Accommodating the instrumental genesis framework within dynamic technological environments. *For the Learning of Mathematics*, *30*(1), 26–31. www.jstor.org/stable/20749435

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education*, *8*, 1270700. https://doi.org/10.3389/feduc.2023.1270700

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, *60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78). Association for Computing Machinery. https://doi.org/10.1145/191666.191703

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, *15*(2), 1–21. https://doi.org/10.1145/3597307

Rabardel, P. (2002). *People and technology – a cognitive approach to contemporary instruments*. https://hal.archives-ouvertes.fr/hal-01020705

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, *9*, 1386075. https://doi.org/10.3389/feduc.2024.1386075

Smith, J. A., & Osborn, M. (2003). Interpretative phenomenological analysis. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 51–80). Sage Publications.

Stangl, W. (2022). *Keyword: 'Adaptivity – online Lexikon für Psychologie und Pädagogik'*. Online Lexicon for Psychology and Education. https://lexikon.stangl.eu/982/adaptivitaet

Suresh, S., Mukherjee, K., Yu, X., Huang, W., Padua, L., & Rogers, T. (2023). Conceptual structure coheres in human cognition but not in large language models. In H.

Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 722–738). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.47

Trouche, L. (2020). Instrumentation in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 404–412). Springer. https://doi.org/10.1007/978-3-030-15789-0_80

Vergnaud, G. (2009). The theory of conceptual fields. *Human Development*, *52*(2), 83–94. https://doi.org/10.1159/000202727

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds., A. R. Luria, M. Lopez-Morillas, M. Cole, & J. V. Wertsch, Trans.). Harvard University Press. (Original manuscripts [ca. 1930–1934]). https://doi.org/10.2307/j.ctvjf9vz4

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Erlbaum. https://doi.org/10.4324/9781315086545

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, 1–21. https://doi.org/10.1007/s11858-024-01629-0

APPENDIX

This section gives a brief description of how GPT-4o was prompted. Figure 8.3 shows the text of the prompts used. While the pre-prompt (first prompt) was the actual start of the conversation, the students did not see it; when ready, they started the conversation by entering a request, as indicated by the researchers (second prompt). This could also include or be followed by an initial question or comment from the student. No fine-tuning or other back-end operations were implemented for this study.

1st prompt (Pre-prompt)	2nd prompt
We want to use ChatGPT to guide 15- to 16-year-old students in conjecturing and proving in arithmetic through the use of algebraic language. You should act as a tutor, without giving direct suggestions or confirmations, but only by asking questions that stimulate students' thinking. At the end, you should invite the student to revise the results obtained in relation to the initial objective. It would be better to ask one question at a time. Each question should be as concise as possible, using language that the user can understand, without losing rigour at the mathematical level.	I have to do a maths problem, this is the text: Consider a natural number. Determine the difference between its square and that of its preceding. Give some numerical examples to help you in the required conjecture. What regularities do you observe? Can you prove what you claim? Don't give me the solution and don't give direct suggestions. Just ask me one question at a time to guide me in solving it. [eventual other request added by the user]

FIGURE 8.3 Prompt and pre-prompt used in the interactions reported.

9

THE POWER OF WELL-DESIGNED MULTIPLE-CHOICE ITEMS TO ENHANCE MATHEMATICS TEACHING AND LEARNING

Bärbel Barzel, Ramona Hagenkötter, Katrin Klingbeil, Fabian Rösken, Anica Stemmer, and Paul Tyrichter

1 Introduction

Multiple-choice (MC) testing is often criticised to focus on what students can remember rather than on the extent of their deep understanding of concepts (e.g. DiBattista & Kurzawa, 2011). However, empirical findings in mathematics education reveal that MC assessments can also assess higher-order thinking skills, such as modelling and problem-solving skills (e.g. Huntley et al., 2009), as well as students' misconceptions (e.g. Klingbeil et al., 2024). The decisive factor here is that the items are well-designed. Therefore, this chapter aims to highlight the power of well-designed MC items for enhancing mathematics teaching and learning. To achieve this, we combine perspectives from educational psychology and mathematics education to identify aspects that determine the quality of MC items for mathematics assessment. First, we summarise scientific findings on why to use MC items and what is known about their potentials and limitations from both perspectives. Next, we provide possible guidelines on how to design and validate MC items, using an example from the digital formative assessment tool SMART (Specific Mathematics Assessments that Reveal Thinking; www.smartvic.com), which we present in more details in Section 3. Finally, we discuss how to utilise MC assessments, especially using digital technology, within mathematics education to enhance both mathematics teaching and learning. We see the potential of well-designed MC items in diagnosing and enhancing students' understanding and in providing teachers with important pedagogical content knowledge about possible misconceptions and relevant facets of a mathematical concept.

DOI: 10.4324/9781003533764-9

2 Scientific findings on potentials and limitations in multiple-choice assessments

In the following, we first clarify relevant terminology before summarising the potentials and limitations on MC assessment from the perspectives of both educational psychology and mathematics education.

MC is a method of assessment where a question is posed and multiple response options are provided, from which the respondent must select the correct one. The main features of MC items are the *stem* and the *response options*. The *stem* (question) serves as the stimulus for the response, presenting a complete idea or concept to ensure the respondent can accurately select the correct answer(s). Following the stem, the *response options* are listed, with one or more correct options and several incorrect options (distractors) (Case & Swanson, 2001). Most commonly used is the single-answer structure, where only one answer is correct (Haladyna, 2004). The response options may be words, phrases, sentences, pictures, or even just numbers. The distractors are designed to trigger cognitive dissonance and to test the knowledge of the respondent. A distractor must appear plausible to test-takers who have not yet mastered the knowledge or skill being assessed. To those who possess the required knowledge, distractors should be clearly incorrect (Haladyna, 2004).

2.1 Educational psychology perspective

MC assessment could be viewed as one of the most persistent and successful educational technologies still widely used in contemporary assessments (Gierl et al., 2017), no longer just in medicine and law degree programmes, as they have found their way into everyday exams in many subjects (Lindner et al., 2015). Research in the field of general educational psychology on MC assessment has demonstrated their potentiality in various educational contexts. Studies indicate that MC tests can reliably measure a range of cognitive skills, from basic recall to higher-order thinking skills, depending on the design of the questions (Fuhrman, 1996). Comparatively, MC questions have advantages over open-ended questions in terms of reliability and efficiency. While open-ended questions can provide deeper insights into student thinking, they require significant time and effort to grade, and their subjective nature can lead to inconsistent scoring. MC questions, on the other hand, offer a high degree of objectivity and consistency in scoring, which is especially crucial for large-scale assessments (Fuhrman, 1996).

2.1.1 Potential in summative and formative assessment

MC questions are most commonly used in the context of summative assessments to evaluate student performance, usually at the end of an instructional period (Stiggins, 2005). The most appropriate uses for MC exams are testing mastery of subject-matter knowledge. The larger the number of students, the more sense it makes to use MC exams for this kind of assessment (Fuhrman, 1996). Therefore, MC testing is appropriate for large-scale assessments, such as standardised national exams, where research-based and well-designed MC questions can comprehensively evaluate student performance. Furthermore, MC tests can also be used in formative assessments, which are designed to provide feedback to teachers or directly to students during learning in order to guide instruction and promote, not merely judge or grade, student success (Stiggins, 2005). Here, they serve this purpose by quickly providing results and identifying student misconceptions.

2.1.2 Potential on the professional development of teachers

Moreover, there are indications that formative MC tests can significantly contribute to the professional development of teachers. MC tests are known for their high validity, ensuring that the constructs they aim to measure are accurately assessed. This reliability provides teachers with confidence that the assessment results reflect the intended learning outcomes. As Heritage (2010, p. 36) states, 'when teachers are making decisions based on assessment results, they need to be sure that the assessment is measuring the construct they think it is measuring'. Engaging with student data from these assessments encourages teachers to reflect on their professional knowledge and instructional practices. For example, programmes that integrate formative assessments into teacher professional development training have shown promising results. Kissi et al. (2023), for instance, found that professional development training with a focus on constructing MC tests improved teachers' competence in developing such assessments, thereby enhancing their ability to use assessment data effectively to guide their teaching.

2.1.3 Potential and limitations regarding language and cultural aspects

Well-designed MC questions can reduce the cognitive load on students by focusing their attention on the key aspects of a problem. This could be particularly helpful for non-native speakers who may struggle with the language but could still engage with the requested content and concepts (Sweller et al., 2011). For example, studies in bilingual education suggest that MC questions can effectively assess and develop understanding among students who are not proficient in the language of instruction. The typical format of MC questions allows for the assessment of students without relying heavily on language skills (Cummins, 2000). Even if MC questions are often seen as a linguistic facilitator (Sweller et al., 2011), the phrasing and the selected contexts and examples can all reflect cultural assumptions that may not be familiar to all test-takers. This is a general problem within test questions (Kim & Zabelina, 2015) but is specifically also discussed in connection with MC issues. This bias can lead to misinterpretation of questions and, subsequently, incorrect answers, not because of a lack of knowledge, but due to cultural differences. Thus, MC questions might not always provide an equitable assessment of all students' abilities, and other, more creative open assessment tasks and formats are required (Kim & Zabelina, 2015).

2.1.4 Limitation of assessable competencies

There is a limitation of competencies testable via MC. Each design of MC items represents a condensation of information. The abundance of possible answers to an open-ended task, such as 'Which story fits the given graph?', is condensed to a few response options in an MC variant. Through this reduction, MC items can only reveal what was previously incorporated into the design of the response options. This means that MC assessments are not suitable for capturing students' creativity or checking their higher-order skills, such as problem-solving, as a whole – but only for checking whether or not certain known thinking patterns or specific cognitive activities are represented. This raises the question of the extent to which MC is suitable for testing process-related competencies, such as problem-solving, which are associated with an act of creativity and openness. However, thoughtfully designed MC items which focus on partial competencies can serve to assess also higher-order thinking (Buckles & Siegfried, 2006; Palmer & Devitt, 2007; cf. Sections 2.2 and 3). But this should never lead to the omission of open-ended assessment tasks, as both formats measure different relevant aspects of comprehension processes (Ozuru et al., 2013), and especially, creativity can only be triggered and made visible via open-ended tasks.

2.1.5 Limitation through the risk of guessing

A further limitation with MC items is the danger that response options are chosen strategically rather than in terms of content. Lindner et al. (2015) report in their review about MC tasks on the university level that a lot of studies mention guessing as a characteristic strategy to solve MC tasks when the respondents have no substantive knowledge of the topic under consideration (Biggs, 1999). However, these results remain doubtful (Zimmerman & Williams, 2003) because little is known about the effects of guessing combined with other sources of error variance determining the reliability of the tests. For example, the issue of guessing often arises in discussions about the appropriate number of response options or the position of the correct answer (Haladyna & Downing, 1993; Haladyna et al., 2002). With regard to the number of response options, it is stated that a high number of response options can increase the reliability of a test and decrease the probability of guessing correctly, but this benefit only materialises if the distractors perform adequately. Often, many distractors are so implausible that they are rarely chosen. Regarding the position of the correct answers, various studies have demonstrated that they are frequently positioned in the middle of the response options, which can lead to strategic rather than content-based answering (Lions et al., 2021). This may be due to the fact that items are easiest when the correct answer is in the first position (Hagenmüller, 2021). To avoid guessing, it is recommended to develop well-designed MC tasks with a randomly selected sequence of the response options and to use a system of crediting that gives less weight to chance than to a substantive response (Downing, 2003; Lindner et al., 2015). These precautions can also mitigate the risk of surface-level learning (Walsh & Seldomridge, 2006), where students merely memorise facts instead of demonstrating deep understanding.

2.2 Mathematics education perspective

With regard to mathematics education, MC questions are also prevalent in assessments due to their ability to assess a wide range of competencies and to uncover misconceptions. These items are particularly effective when the distractors are designed based on typical misconceptions, as evidenced by numerous studies. For example, misconceptions related to irrational numbers (Kidron, 2018), limits and continuity (Bezuidenhout, 2001), infinity (Kolar & Cadez, 2012), and derivation (Orton, 1983) have been effectively investigated – an important basis for a careful design of distractors (Rach & Ufer, 2020).

2.2.1 Assessment of mathematical competencies

Huntley et al. (2009) demonstrated that MC items can assess not only technical procedural skills, like basic manipulations and calculations, but also complex cognitive processes, such as translating words into mathematical symbols (i.e. modelling) and identifying and applying mathematical methods to solve problems. For summative assessment, research has shown the effectiveness of MC questions in testing at the transition from school to university. For instance, Besser et al. (2021) developed a reliable and valid entry test comprising 84 highly objective MC tasks to assess secondary education–related mathematical knowledge for university entrants. Similarly, Rach and Ufer (2020) utilised MC tasks to assess not only procedural or declarative knowledge but also conceptual knowledge and higher-order skills. In formative assessment, Kolar and Cadez (2012) used MC items to capture intuitive knowledge about infinity and directly prompt whole-classroom discussions.

Meanwhile, Stacey et al. (2018) employed MC items in the SMART assessment system for various mathematical topics to strengthen teachers' knowledge and enhance individual formative feedback. The use of MC tasks during the learning process has also been explored. Große (2017), for example, investigated whether MC tasks could foster translation and interpretation skills in solving modelling problems related to linear systems of equations. Using a two-factorial design with four experimental conditions (i.e. translation choice and interpretation choice, both with and without MC), she found no clear benefit of using MC tasks to enhance learning within word problems. This suggests that while MC questions can effectively diagnose and assess, their role in fostering deep learning and interpretation may be limited.

2.2.2 Language and accessibility

In mathematics, MC questions often include diagrams, symbols, and contextual clues, which can help bridge language gaps and make content more accessible to non-native speakers (Abedi & Lord, 2001). This is particularly important in diverse classrooms, where language barriers might otherwise hinder the accurate assessment of mathematical understanding.

In almost all findings on potentials and limitations from both the perspective of educational psychology and mathematics education, a good design of MC questions is emphasised. In the following, we will crystallise practical aspects that can guide designing and validating high-quality MC questions.

3 Designing and validating high-quality MC items for mathematics education

In the literature, several (rather general) guidelines on designing MC items can be found (e.g. Haladyna et al., 2002). However, by integrating perspectives from educational psychology and mathematics education, we aim to provide possible guidelines with the focus on MC items within mathematics teaching and learning. We illustrate our proposed procedure using an example from SMART (Specific Mathematics Assessments that Reveal Thinking), a digital formative assessment tool in mathematics education based on MC items, specifically the SMART test 'Meaning of Letters' (Klingbeil et al., 2024). Based on an analysis of students' response patterns, the tool provides teachers with feedback on each student's level of conceptual understanding, possible misconceptions, or common errors, along with didactical background information on the tested topic and fitting recommendations for further instruction (Stacey et al., 2018). Our proposed approach is not definitive, but we believe it offers a promising and suitable method to design and validate high-quality MC items for mathematics education.

3.1 Design

Here, we focus on four rather general guidelines from a list of 31 supposed guidelines from Haladyna et al. (2002). While most of the 31 guidelines refer to more general matters as wording, clearness, and layout, these four explicitly refer to the content and the distractors. Therefore, they are of specific relevance for mathematics education and need to be addressed in a subject-specific manner:

1. Every item should reflect specific content and a single specific mental behavior. . . .
2. Base each item on important content to learn; avoid trivial content. . . .

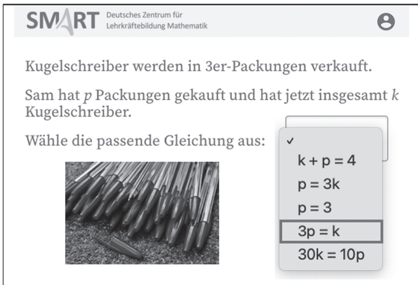
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.

(Haladyna et al., 2002, p. 312)

The first two guidelines address content concerns, while the other two relate to writing the response options. Accordingly, in the following, we will first discuss the content of well-designed MC items and then the writing of distractors.

3.1.1 Content of MC items

To ensure that each item reflects specific content and a single specific mental behaviour, as well as to base each item on important content to learn (i.e. first and second guideline from Haladyna et al., 2002), the learning goals that ought to be assessed must first be clearly determined. Moreover, to be able to gain deep insights into students' thinking, these learning goals need to be broken down into partial competencies and specific *Grundvorstellungen* (vom Hofe, 1995) that students need to develop. It is important not only to include procedural skills and declarative knowledge but also to especially address the understanding of mathematical concepts and operations by specifying modes and facets of knowledge (see, for example, Barzel et al., 2013). For example, when students are supposed to learn about algebraic expressions, it is not enough to know about transformation rules and conventional terms; they especially need to know about the meaning and connections of algebraic expressions. They have to understand the structure of algebraic expressions and be able to express a situational relationship with algebra. For this, it is essential to understand the meaning of algebraic letters – such as a generalised number, an unknown number, or a varying quantity (Arcavi et al., 2016). With this in mind, our example of the SMART test 'Meaning of Letters' focuses on students' basic understanding of algebraic notation, including the meaning students assign to the letters of the alphabet that are used to write



Proposed thinking:
(Note: Instead of the letter *b*, in German, *k* is used since 'Kugelschreiber' is the translation of 'biro'.)

One biro plus a pack of biros is 4 altogether.
A pack contains 3 biros.
A pack has 3.
3 times the number of packs equals the number of biros.
Sam bought 10 packs and has 30 biros now.

FIGURE 9.1 Example from the SMART test 'Meaning of Letters'.

Source: Klingbeil et al. (2024, p. 717).

algebra. Specifically, the test has been developed to assess whether students understand that letters in algebra represent numbers.

3.1.2 Writing the distractors

Subsequently, when writing the distractors, following the relevant guidelines from Haladyna et al. (2002), it is necessary to make all distractors plausible and to use typical student errors and common misconceptions. This can be achieved by drawing on research literature, student answers to open-ended or constructed-response items, or think-aloud student interviews (e.g. Gierl et al., 2017). Accordingly, the distractors of the SMART test 'Meaning of Letters' are based on a wide range of elementary algebra items in open-ended pen-and-paper format as well as clinical interviews identifying common student misconceptions, which, 'in contrast to careless errors, . . . lead to predictable errors in student work' (Akhtar & Steinle, 2013, p. 36). For instance, the example MC item (see Figure 9.1) focuses especially on the letter-as-object misconception, where the algebraic letter is thought of not as a number but as a reference to an object or an abbreviation of its name.

3.2 Validation

After developing the MC items, their quality must be evaluated. To do so, various approaches are available in the literature. For example, from a psychological perspective, Moreno et al. (2015) identify three properties of validity of MC items: representativeness, precision, and differentiation. *Representativeness* ensures completeness and parsimony of relevant elements. *Precision* refers to the clarity and unambiguity of each element, proven by consistent understanding and use. *Differentiation* ensures that elements understood as different and independent are mutually exclusive. Another approach involves analysing the responses that examinees give regarding difficulty, discriminatory power, and effectiveness of the distractors (e.g. DiBattista & Kurzawa, 2011). The *difficulty* index describes the proportion of examinees selecting the correct option, *discriminatory power* indicates the likelihood of more knowledgeable students selecting the correct option compared to less knowledgeable ones, and distractor *effectiveness* is demonstrated by at least some examinees selecting it, with fewer knowledgeable students doing so.

However, if MC items and distractors are designed with the purpose not only to distract but also to elicit further diagnostic information to enhance mathematics teaching and learning, these approaches may not suffice. For example, with regard to formative assessment in particular, Gikandi et al. (2011, p. 2337) argue

> that it is necessary to reconceptualise and redefine validity and reliability within the context of formative assessment because the typical definitions applied in summative assessment are limited to quantitative conceptualizations, which is not sufficient to establish validity and reliability within the contest of formative assessment. . . . Therefore, a qualitative or mixed methods approach is often required to establish the degree of validity and reliability in formative assessment.

One possibility for such a qualitative or mixed-methods approach is the validation framework for formative assessment proposed by Hopster-den Otter et al. (2019). Within their framework, the authors emphasise the importance of alignment with the teaching and learning process, the need for fine-grained information, and especially the relevance of the used facet of formative assessment. For the validation, the authors recommend to 'build and evaluate an argument that helps test developers demonstrate that assessment scores are sufficiently useful for their intended purpose' (p. 719). Following these recommendations, we used a qualitative approach to validate the MC items of the SMART test 'Meaning of Letters' (Klingbeil et al., 2024). Specifically, we conducted cognitive interviews with a small sample of students and found that the chosen response options were predominantly in line with students' explanations, particularly regarding the letter-as-object misconception, and that the investigated misconceptions were not evoked by the response options. Additionally, we analysed written explanations of 600 seventh and eighth graders to the example item (see Figure 9.1) and found that the students primarily chose response options for the reasons initially assumed when designing them. Thus, we were able to ensure that the distractors are plausible and represent typical students' errors (cf. guidelines 29 and 30 by Haladyna et al. (2002) in Section 3.1).

4 Utilisation of MC items in mathematics education

To describe the ways MC items might be used to enhance mathematics teaching and learning, it is important to differentiate between summative and formative assessment. For both forms of assessment, teachers can use different scenarios, which will be presented in the following considering the role of technology, the moment of use in a lesson series, the agent in the process, and the enhancement of mathematics teaching and learning.

4.1 Summative assessment

With regard to summative assessments, two main scenarios can be distinguished: in-class tests and large-scale assessments.

Currently, summative assessments in both scenarios are predominantly realised in pen-and-paper format. Nonetheless, digital technology is becoming more and more introduced in large-scale assessments in mathematics education (cf. Chapters 3 and 4 of this book), and benefits from MC items are used. First attempts in implementing MC items in summative assessment through a student response system, such as, for example, the study by Premkumar (2016), indicate that instructors at university level experience such utilisation as less time-consuming, more efficient, and secure compared to pen-and-paper. Furthermore, the results reveal that students perceive this form of assessment as engaging and satisfying, primarily because immediate feedback was provided.

Summative assessment is predominantly realised at the end of a lesson series or a school year, such as (centralised) final examinations. Looking at the role of students and teachers in these scenarios, students are rarely involved in the construction or evaluation of tasks. Thus, teachers are often the main agents in constructing and evaluating summative assessment tasks. The vision of involving learners in the construction of test items by Nieminen et al. (2024) represents a notable step forward in rethinking assessment practices. Earlier examples, such as Rapke (2016), have also explored collaborative processes between teachers and students in exam development. Regarding large-scale assessments, sometimes even an external agent is involved in the assessment process, so neither teachers nor students play an active role (Adie et al., 2018). The role of teachers, and especially students, can be strengthened when at least the results of a summative assessment are used formatively. This might include discussing typical errors or reworking particular tasks.

4.2 Formative assessment

In contrast, the utilisation of MC items in formative assessment may be realised very differently. In the following, we will present three exemplary scenarios. Of course, similar scenarios realised in pen-and-paper assessments are also possible, but as we will see in the following, digital technology and, particularly, the use of MC items, offer benefits, such as a quick evaluation.

4.2.1 Scenario 1: Digital diagnostic tests

A first scenario is the use of MC items in digital tests to diagnose different aspects of students' learning, such as procedural skills or conceptual knowledge. At this point, it is the design of the specific MC items that determines the diagnostic focus and the information to teachers. The range goes from quotas of right answers to information about stages of understanding arising from an analysis of students' response patterns (e.g. SMART, cf. Section 3). Moreover, digital diagnostic tools might provide teachers with didactical background information concerning the tested topic or fitting teaching suggestions to the diagnostic outcome (e.g. SMART). In this scenario, technologies using MC items play a crucial role in supporting teachers in the assessment process through an easy conduction and automatic evaluation, making the process less time-consuming (Drijvers, 2018). MC items in formative assessment can be used at different moments of a lesson series, for example, at the beginning, to check necessary prior knowledge; at the end, to prepare students for an in-class test; or in between, to diagnose students' strengths and weaknesses and adapt further teaching (Cusi et al., 2024). Furthermore, in this scenario, students are not actively engaged in the construction or evaluation of the MC items, but they benefit from lessons that are more precisely adapted to their individual needs. Even if teachers are not actively involved in the design of tests, if they use ready-made ones (e.g. SMART), they, too, can benefit from them. Research found that '[e]xamining a particular question and identifying the distracter the student has selected can provide teachers with an insight into the students' misconceptions about a particular concept' (Rogers & Zoumboulis, 2015, p. 117). This indicates that MC items might also contribute to enhancing or refreshing teachers' pedagogical content knowledge with respect to the capabilities of their own class, thereby improving their teaching not only for the current moment but also for future lessons. The well-designed MC items in SMART, for instance, can allow such improvement.

4.2.2 Scenario 2: Classroom response systems to foster whole-class discussions

The second scenario shows the utilisation of MC items using a classroom response system (CRS) to engage students in whole-class discussions. That means classroom situations where students answer MC items using a CRS that immediately displays an overview of all students' answers (see, for example, Cusi et al., 2024). This forms the basis for adapting the learning process, as the teacher can determine instantly how to deal with the gained information, by, for example, directly orchestrating a peer or whole-class discussion (Kay & LeSage, 2009). Here, the power of MC items lies in the fact that, on the one hand, they allow for a quick response and, at the same time, offer different aspects of discussion that are addressed in the response options. In this scenario, MC items might be used in various moments of a lesson series. For instance, this form of classroom discussions can help reactivate necessary prior knowledge at the beginning of a lesson series, repeat or strengthen currently treated content shortly before an in-class test, or promote further learning in the middle of a lesson series. Students and teachers are both highly engaged in these discussions, but in classroom activities outside research situations, it is normally the teacher who designs the MC items, making the teacher's role more active in this regard. This activity can be realised differently, as teachers can either select or adapt given MC items (e.g. from existing diagnostic tools) or create new ones themselves (Cusi et al., 2024). If they do so, it is important to respect guidelines for the design of high-quality MC items in order to gain impactful information and thus raise fruitful discussions (cf. Section 3). For instance, Gustafsson (2023) analysed whole-class discussions in mathematics classrooms with regard to teacher and student actions after the implementation of a CRS using MC items focusing on students' conceptual understanding, understanding of procedures, as well as common misconceptions and mistakes. He found that although the teacher dominated the discussion, students had opportunities to actively engage in mathematical conversation. Despite slight differences in the students' and teachers' actions depending on the focus of the discussed MC item, he concludes (p. 880):

> The findings in this study indicate that if teachers wisely implement well-designed MC tasks supported by CRS, and apply productive teacher actions during whole-class discussions, they can achieve productive discussions in which students have the opportunity to improve their learning.

In addition, in a literature review by Kay and LeSage (2009) including different forms of the use of CRS with MC items, benefits were found concerning the classroom environment (i.e. students' attention, participation, and engagement), students' learning (i.e. interaction, discussion, learning performance, and quality of learning), and the assessment process (i.e. feedback and formative assessment practices).

4.2.3 Scenario 3: Digital MC tests to evoke peer and self-assessment

In the third scenario, MC items are used in an online test where a fixed set of questions is proposed to all students of a course. Students have the possibility to conduct this test as often as they wish in a given time slot, as well as to save and revise their answers each time they access the test. After the access period has ended, or even after each saved attempt, students receive feedback on the correctness of the answers in their latest submission (e.g. Roberts, 2006). Such a scenario might be realised primarily at the end or in the middle of a lesson series to gain information about students' efforts and to provide students the opportunity to improve during the test attempts by working a long time on the same questions. In this scenario, students have an active role in conducting and regulating the assessment process, even though they are not involved in item construction or evaluation. The teachers' role is more or less passive during the assessment process, as they do not interact with the students. It is the teachers' job to create the MC items and to determine the consequences of the final results. This arrangement of MC items in a test, which students may work on multiple times, allows extra learning for students by giving them the opportunity to identify where difficulties lie and to work out the correct answers in order to improve their learning (Roberts, 2006). In this regard, such a scenario allows for a form of self-assessment. The aspect of self-assessment might be strengthened when feedback is given not only at the end of the access period but also after each submitted attempt (Andrade & Valtcheva, 2009). Furthermore, the fact that all students in a course get the same items might encourage discussions with peers during and after the test. Concerning the role of the teacher, information about which items are attempted frequently can provide hints about aspects that may benefit from further instruction or a greater emphasis in upcoming lessons (Roberts, 2006). Despite the benefit of an immediate evaluation, MC items in this scenario potentially also facilitate the process of self-assessment as students can more easily compare their ideas and argumentations for specific response options than in answers to open-ended questions.

4.2.4 Students as active agents in the assessment process

In all the studies discussed in the three scenarios, it was the researcher who initially designed the MC items. By designing and implementing MC items themselves, teachers could also take a more active role in the entire process. If they do so, in the described settings, it is mostly the teacher who is the main agent in the formative assessment practice even though the engagement of students increases from one scenario to the other. However, students might also be the main agents in a formative assessment practice using MC items. For example, it is possible to involve students in the design of MC items that will afterwards be discussed using a classroom response system (scenario 2) or a digital tool for self-assessment (scenario 3). Students might also take an active role in evaluating solutions of tasks and giving feedback to their peers. Thus, a form of peer instruction can be realised that might enhance mathematics learning. Such a combination of particular aspects of each scenario might increase the active involvement, and thus the potential learning of students, because actively involving students in the creation and evaluation of MC items during a formative assessment highly enlarges their activity and agency in the assessment process. This approach transforms students from passive participants into active contributors and provides teachers with deeper insights into their thinking (e.g. Jones, 2019; Ribosa & Duran, 2022).

5 Conclusion

By integrating perspectives from educational psychology and mathematics education, the present chapter illustrates how MC items can be designed and validated to assess students' understanding and identify possible misconceptions so that mathematics teaching and learning may be enhanced. We followed current recommendations from Gikandi et al. (2011) and Hopster-den Otter et al. (2019), distinguishing ourselves from the criteria of validity from those in (educational) psychology (see, for example, DiBattista & Kurzawa, 2011; Moreno et al., 2015) by focusing on a more qualitative validity within formative assessment. This involves the importance of the alignment with the teaching and learning process, the need for fine-grained information, and especially the relevance of the use. This chapter further emphasises that MC items are utilised in various ways in both formative and summative assessments. They serve to diagnose and enhance students' understanding and provide teachers with valuable insights into possible misconceptions and relevant facets of a mathematical concept, thereby enhancing their pedagogical content knowledge. Thus, the power of well-designed MC items for improving mathematics teaching and learning appears to be underestimated. More research is needed to understand their specific role in mathematics education. Notably, actively involving students in the creation and evaluation of MC assessments (whether summative or formative) seems to be a promising opportunity for fostering mathematics teaching and learning.

Reference list

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Adie, L. E., Willis, J., & Van der Kleij, F. M. (2018). Diverse perspectives on student agency in classroom assessment. *Australian Educational Researcher*, *45*, 1–12. https://doi.org/10.1007/s13384-018-0262-2

Akhtar, Z., & Steinle, V. (2013). Probing students' numerical misconceptions in school algebra. In V. Steinle, L. Ball, & C. Bardini (Eds.), *Mathematics education: Yesterday, today and tomorrow. Proceedings of the 36th annual conference of the mathematics education research group of Australasia* (pp. 36–43). Mathematics Education Research Group of Australasia Inc.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, *48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Arcavi, A., Drijvers, P., & Stacey, K. (2016). *The learning and teaching of Algebra: Ideas, insights and activities* (1st ed.). Routledge. https://doi.org/10.4324/9781315545189

Barzel, B., Leuders, T., Prediger, S., & Hußmann, S. (2013). Designing tasks for engaging students in active knowledge organization. In C. Margolinas (Ed.), *Task design in mathematics education. Proceedings of ICMI study 22* (pp. 283–292). Oxford.

Besser, M., Göller, R., Ehmke, T., Leiss, D., & Hagena, M. (2021). Entwicklung eines fachspezifischen Kenntnistests zur Erfassung mathematischen Vorwissens von Bewerberinnen und Bewerbern auf ein Mathematik-Lehramtsstudium. *Journal für Mathematik-Didaktik*, *42*, 335–365. https://doi.org/10.1007/s13138-020-00176-x

Bezuidenhout, J. (2001). Limits and continuity: Some conceptions of first-year students. *International Journal of Mathematics Education in Science and Technology*, *32*(4), 487–500.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, *18*, 57–75. https://doi.org/10.1080/07294360.2012.642839

Buckles, S., & Siegfried, J. J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education*, *37*(1), 48–57. https://doi.org/10.3200/JECE.37.1.48-57

Case, S. M., & Swanson, D. B. (2001). *Constructing written test questions for the basic and clinical sciences*. National Board of Examiners.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Multilingual Matters. https://doi.org/10.21832/9781853596773

Cusi, A., Aldon, G., Barzel, B., & Olsher, S. (2024). Rethinking teachers' formative assessment practices within technology-enhanced classrooms. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 1–31). Springer International Handbooks of Education.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2). https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et évaluation en éducation*, *41*(1), 41–66. http://dx.doi.org/10.7202/1055896ar

Fuhrman, M. (1996). Developing good multiple-choice tests and test questions. *Journal of Geoscience Education*, *44*(4), 379–384. https://doi.org/10.5408/1089-9995-44.4.379

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Große, C. S. (2017). Effects of multiple choice options in mathematics learning. *European Journal of Science and Mathematics Education*, *5*(2), 165–177.

Gustafsson, P. (2023). Productive mathematical whole-class discussions: A mixed-method approach exploring the potential of multiple-choice tasks supported by a classroom response system. *International Journal of Science and Mathematics Education*, *22*(4), 861–884. https://doi.org/10.1007/s10763-023-10402-w

Hagenmüller, B. (2021). On the impact of the response options' position on item difficulty in multiple-choice-items. *European Journal of Psychological Assessment*, *37*(4), 290–299. https://doi.org/10.1027/1015-5759/a000615

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge/Taylor & Francis Group.
Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press. https://doi.org/10.4135/9781452219493
Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*, *56*(4), 715–732. https://doi.org/10.1111/jedm.12234
Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics? *Pythagoras*, *69*, 3–16. https://doi.org/10.4102/pythagoras.v0i69.41
Jones, J. A. (2019). Scaffolding self-regulated learning through student-generated quizzes. *Active Learning in Higher Education*, *20*(2), 115–126. https://doi.org/10.1177/1469787417735610
Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education*, *53*(3), 819–827. https://doi.org/10.1016/j.compedu.2009.05.001
Kidron, I. (2018). Students' conceptions of irrational numbers. *International Journal of Research in Undergraduate Mathematics Education*, *4*, 94–118. https://doi.org/10.1007/s40753-018-0071-z
Kim, K. H., & Zabelina, D. (2015). Cultural bias in assessment: Can creativity assessment help? *The International Journal of Critical Pedagogy*, *6*(2), 129–148.
Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education*, *8*, 1154592. https://doi.org/10.3389/feduc.2023.1154592
Klingbeil, K., Rösken, F., Barzel, B., Schacht, F., Stacey, K., Steinle, V., & Thurm, D. (2024). Validity of multiple-choice digital formative assessment for assessing students' (mis)conceptions: Evidence from a mixed-methods study in algebra. *ZDM Mathematics Education*, *56*, 713–726. https://doi.org/10.1007/s11858-024-01556-0
Kolar, V. M., & Cadez, T. H. (2012). Analysis of factors influencing the understanding of the concept of infinity. *Educational Studies in Mathematics*, *80*, 389–412. https://doi.org/10.1007/s10649-011-9357-7
Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung. *Zeitschrift für Pädagogische Psychologie*, *29*(3–4), 133–149. https://doi.org/10.1024/1010-0652/a000156
Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdove, N., Jiménez, D., Blanco, M., Ortega, G., & Lemarié, J. (2021). The position of distractors in multiple-choice test items: The strongest precede the weakest. *Frontiers in Education*, *6*, 731763. https://doi.org/10.3389/feduc.2021.731763
Moreno, R., Martínez, R. J., & Muniz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, *27*(4), 388–394. https://doi.org/10.7334/psicothema2015.110
Nieminen, J. H., Haataja, E., & Cobb, P. (2024). From active learners to knowledge contributors: Authentic assessment as a catalyst for students' epistemic agency. *Teaching in Higher Education*, 1–21. https://doi.org/10.1080/13562517.2024.2332252

Orton, A. (1983). Students' understanding of differentiation. *Educational Studies in Mathematics*, *14*, 235–250. https://doi.org/10.1007/BF00410540

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, *67*(3), 215–227. https://doi.org/10.1037/a0032918

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. *BMC Medical Education*, *7*, 49. https://doi.org/10.1186/1472-6920-7-49

Premkumar, K. (2016). Use of student response systems for summative assessments. *Creative Education*, *7*, 1851–1860.

Rach, S., & Ufer, S. (2020). Which prior mathematical knowledge is necessary for study success in the university study entrance phase? Results on a new model of knowledge levels based on a reanalysis of data from existing studies. *International Journal of Research in Undergraduate Mathematics Education*, *6*, 375–403. https://doi.org/10.1007/s40753-020-00112-x

Rapke, T. (2016). A process of students and their instructor developing a final closed-book mathematics exam. *Research in Mathematics Education*, *18*(1), 27–42. https://doi.org/10.1080/14794802.2015.1134342

Ribosa, J., & Duran, D. (2022). Do students learn what they teach when generating teaching materials for others? A meta-analysis through the lens of learning by teaching. *Educational Research Review*, *37*, 100475. https://doi.org/10.1016/j.edurev.2022.100475

Roberts, T. S. (2006). The use of multiple choice tests for formative and summative assessment. In *Proceedings of the 8th Australasian computing education conference (ACE2006)* (pp. 175–180). Australian Computer Society, ACM Digital Library. https://hdl.handle.net/10018/6995

Rogers, P., & Zoumboulis, S. (2015). Using multiple choice questions to identify and address misconceptions in the mathematics classroom. In N. Davis, K. Manuel, & T. Spencer (Eds.), *Mathematics: Learn, lead, link. Proceedings of the 25th biennial conference of the Australian Association of Mathematics Teachers Inc*. (pp. 112–120). The Australian Association of Mathematics Teachers (AAMT) Inc. https://doi.org/10.13140/RG.2.1.1564.5925

Stacey, K., Steinle, V., Price, B., & Gvozdenko, E. (2018). Specific mathematics assessments that reveal thinking: An online tool to build teachers' diagnostic competence and support teaching. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers* (pp. 241–261). Springer International Publishing. https://doi.org/10.1007/978-3-319-66327-2_13

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *The Phi Delta Kappan*, *87*(4), 324–328. https://doi.org/10.1177/003172170508700414

Sweller, J., Ayres, P. L., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

vom Hofe, R. (1995). *Grundvorstellungen mathematischer Inhalte*. Spektrum.

Walsh, C., & Seldomridge, L. (2006). Critical thinking: Back to square two. *The Journal of Nursing Education*, *45*(6), 212–219. https://doi.org/10.3928/01484834-20060601-05

Zimmerman, D., & Williams, R. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357–371. https://doi.org/10.1177/0146621603254799

10

EXAMPLE-GENERATION TASKS FOR COMPUTER-AIDED ASSESSMENT

Maria Fahlgren and Mats Brunström

1 Introduction

Today, the integration of computer-aided assessment (CAA) systems into mathematics education is extensive. However, scholars in the realm of technology-enhanced assessment caution against the tendency for such assessment to prioritise basic mathematical skills and to focus on the correctness of final answers (Hoogland & Tout, 2018). This tendency arises because tasks and feedback of this nature are the most straightforward to implement in CAA systems. In this chapter, we address this issue by focusing on a specific type of task: example-generation tasks. In such tasks, students are prompted to create examples that meet certain conditions (Watson & Mason, 2005). Researchers advocate for example-generation tasks as a means to actively engage students in developing a deeper understanding of mathematics (Antonini et al., 2011; Bills et al., 2006). To generate examples, students often need to be creative and devise strategies based on their conceptual understanding. Because example-generation tasks can often be automatically assessed, they are well-suited for integration into CAA systems (Kinnear et al., 2022; Yerushalmy et al., 2017).

Guided by the theory of example spaces proposed by Watson and Mason (2005), we elaborate on various types of example-generation tasks suitable for the combined use of a dynamic mathematics software (DMS) environment and a CAA system. In light of findings from our previous research, we discuss how tasks and associated feedback could be designed to enrich students' example spaces and address key ideas targeted by a task. We compare two types of example-generation tasks and reflect on how the task type, as well

DOI: 10.4324/9781003533764-10

as the addition of adaptive feedback provided through the CAA system, might affect student performance.

1.1 Digital feedback

CAA systems facilitate instant feedback. However, this feedback has primarily centred on the correctness of the final answers due to its straightforward implementation in CAA systems (Rønning, 2017). In contrast, Sangwin (2013), the developer of the established CAA system STACK, takes a different stance. STACK is primarily designed for formative feedback rather than examination purposes. As a result, STACK can be used to not only assess answer correctness but also provide elaborated feedback adapted to specific student responses. Such feedback can include explanations for why a particular response is incorrect, as well as offer conceptual hints or guidance for solving the task (Sangwin, 2013).

To further increase the learning potential when using CAA systems, researchers propose integrating another type of technology: DMS environments (e.g. Yerushalmy et al., 2017). This type of technology is widely recognised as a tool for promoting inquiry and fostering students' conceptual understanding in mathematics. It is the instant feedback on students' action that makes it possible for them to use DMS environments to explore, conjecture, verify, and have reflective discussions. Even if DMS feedback does not explicitly provide hints on how to proceed, it offers information that could be used in a productive way by the user (Moreno-Armella et al., 2008; Olsson, 2018).

1.2 Example spaces

Watson and Mason (2005) suggest prompting students to generate examples that fulfil certain conditions as a pedagogical approach in the teaching and learning of mathematics. They define 'example spaces' as collections of examples that satisfy specified conditions, and they distinguish between personal, conventional, and collective and situated example spaces. *Personal* example spaces are the range of examples an individual can come up with, whereas *conventional* example spaces comprise examples generally understood by mathematicians and commonly found in textbooks. *Collective* and *situated* example spaces refer to the examples shared by a group at a specific time (Watson & Mason, 2005).

In their elaboration on example spaces, Watson and Mason (2005) introduce two fundamental concepts: dimensions of possible variation and ranges of permissible change. These concepts stem from the theory of variation, particularly from Marton and colleagues' notion of dimensions of variation (e.g. Marton & Booth, 1997). *Dimensions of possible variation* (DofPV) represent the attributes of an example that can be adjusted without compromising its defining characteristics. The associated *ranges of permissible change* (RofPCh) delineate the extent to which these dimensions can be modified while still ensuring the example's validity (Watson & Mason, 2005). For instance, if the task is to generate examples of quadratic functions, $f(x) = ax^2 + bx + c$, with a specific *y*-intercept, both parameters *a* and *b* are DofPV. Their acceptable values, that is, all real numbers except zero for *a* and all real numbers for *b*, correspond to the associated RofPCh.

The richness of students' personal example spaces can serve as an indicator of their mathematical understanding (Watson & Mason, 2005; Zazkis & Leikin, 2007). To prompt students to enrich their existing example spaces, Watson and Mason (2005) propose various types of example-generation tasks.

1.3 Example-generation tasks

This chapter focuses on two particular types of example-generation tasks. In the first type of task (type 1), students are requested several examples that fulfil the same conditions. This approach encourages students to explore the DofPV and its associated RofPCh (Mason, 2011). To further enhance the opportunities for students to generate examples beyond their initial thoughts, researchers recommend asking for examples that differ as much as possible (Watson & Mason, 2005; Zaslavsky & Zodik, 2014). For instance, Zaslavsky and Zodik (2014) encouraged 'notable variation' between examples, pushing students beyond familiar and prototypical examples toward more sophisticated ones. Achieving this requires students to compare examples, identifying similarities and differences.

In the other type of example-generation task (type 2), additional constraints are progressively added to the initial condition. According to Watson and Mason (2005), this addition of constraints often leads to new avenues of exploration and fosters creativity among learners. When a new constraint renders typical student examples invalid, they are challenged to devise innovative solutions to meet the additional criteria.

Although example-generation tasks have been proposed in the literature, they are time-consuming to correct manually because such tasks often permit multiple correct responses. This problem is particularly relevant at the university level, where there are often large teaching groups. Consequently, researchers in the field of digital assessment in mathematics emphasise the usefulness of employing CAA systems for automated corrections of these types of task (Kinnear et al., 2022). Furthermore, researchers highlight the potential to provide elaborated and adapted feedback to students working with example-generation tasks embedded in a CAA system. In STACK, for instance, it is possible to compare properties of different examples provided by a student and give feedback based on this comparison. This is achieved by using 'formative potential response trees', as described in STACK Docs (n.d.). Kinnear and Kontorovich (2024) introduce the notion 'interactive example-generation tasks' in their elaboration on tasks 'that prompt students for further examples based on the examples they have given so far' (p. 1).

2 Research context

In this chapter, we draw on data collected as part of a research and development project. The overarching aim of this project was to understand how a carefully designed combination of two specific types of digital technology, CAA and DMS, can be utilised to enhance students' engagement and conceptual understanding in mathematics. The research took place at a Swedish university during the autumns of 2020, 2021, and 2022, involving three groups of first year engineering students enrolled in an introductory calculus course. The cohorts consisted of 256 students in 2020, 235 students in 2021, and 205 students in 2022. As part of their assignment, students were engaged in small group activities that involved tasks that required a joint group answer as well as tasks that called for an individual response. These tasks were embedded in a CAA system (in this case, Möbius[1]). Several of the tasks tried out in the project are example-generation tasks. In the following, we will elaborate on three of these example-generation tasks, two of the first type (type 1) and one of the second type (type 2). In all three tasks, students are advised to check their responses using a DMS (in this case, GeoGebra[2]) before entering them into the CAA system. This way, they receive immediate feedback that can be used to revise responses that do not meet all the given conditions.

3 Task redesign to enrich students' example spaces

What follows is a detailed description of each of the three tasks. Additionally, findings from the different cohorts will be introduced for each task, expressed in terms of the collective and the situated example spaces. For each task, we will also elaborate on how the task could be redesigned to enrich students' example spaces by promoting the activation of further DofPV. More precisely, we will elaborate on the opportunity of changing the type of example-generation task and how the provision of adapted feedback could further improve the redesign. Since the students are encouraged to verify their answers in the DMS environment, we chose to focus the discussion on CAA feedback for correct answers.

3.1 Task 1

3.1.1 The original task

The original version of task 1 is an example of the first type of example-generation task (type 1). In this task (see Figure 10.1), the students are asked to provide two examples of functions with specified asymptotes:

two vertical asymptotes and one horizontal asymptote. This is an individual task, where students receive different numerical values of the asymptotes. The main key idea addressed in this task is the relationship between asymptotes (in a graph) and the corresponding function formula. Students are encouraged to use a DMS to verify their proposed functions before submitting their responses to the CAA system. This approach enables them to receive immediate feedback through the graphical representations of the functions.

> Give examples of two different functions, *f* and *g*, both of which have
>
> - two vertical asymptotes, $x = 3$ and $x = -6$ as well as
> - a horizontal asymptote, $y = 2$.
>
> Note:
>
> - Group members may have received different asymptotes.
> - Check in GeoGebra if your suggested functions really have the given asymptotes.

FIGURE 10.1 Original version of task 1.

Since there are various function types adaptable to the conditions, we regard the type of function as one DofPV. However, given a previous task focusing on rational functions, we anticipate students to employ this function type. Additionally, for a rational function to meet the specified conditions, the degrees of its numerator and denominator can vary but must be equal when represented as a single quotient. Thus, we propose this as an additional DofPV, with all integers greater than one as the associated RofPCh. Most likely, students will offer rational functions with numerators and denominators of degree 2.

Furthermore, a rational function can be represented in multiple formats. In a previous study (Brunström et al., 2022), we identified three primary formats for expressing the function formula in student responses, as depicted in Table 10.1.

TABLE 10.1 Three main ways of expressing the function formula as a response to task 1

	Function formula format	*Example*
A	Single quotient	$f(x) = \dfrac{2x^2 + ax + b}{(x+6)(x-3)}$
B	Partial fraction, reduced quotients, and a constant term (i.e. the horizontal asymptote)	$f(x) = \dfrac{a}{x+6} + \dfrac{b}{x-3} + 2$
C	Reduced quotient and a constant term (i.e. the horizontal asymptote)	$f(x) = \dfrac{ax+b}{(x+6)(x-3)} + 2$

Even though these are just different ways to express rational functions, we consider the choice of function formula format as a DofPV. It is worth noting that we employ the same letters, *a* and *b*, to denote the parameters that can be varied, albeit they may influence the function differently, depending on the formula format (A, B, or C). As both parameter *a* and parameter *b* can have various values, we consider the parameter choice as one DofPV. Regarding the associated RofPCh, the values of parameters *a* and *b* could be any real number (except zero in certain function formula formats).

To summarise, the recognised conventional example space encompasses four DofPVs that students could utilise in constructing the second example: the function type (DofPV1), the numerator and denominator degrees (DofPV2), the function formula format (DofPV3), and the parameter values *a* and *b* (DofPV4).

3.1.2 The collective and situated example spaces observed

Data consisting of 491 student responses were collected in 2020 and 2021 (Fahlgren & Brunström, 2023). In total, 479 students (out of 491) gave a first example, among which 465 were correct. In Table 10.2, the number of students that activated the different DofPVs to construct the second example is summarised.

TABLE 10.2 Activated DofPVs when constructing the second example

	DofPV1	*DofPV2*	*DofPV3*	*DofPV4*
Number of students	0	20	34	397

Predominantly, students changed the parameter *a* and/or the parameter *b* (DofPV4) to construct their second example. Since one of the preceding tasks concerns a rational function, it is not surprising that all students responded with rational functions and, thus, did not activate DofPV1. Moreover, most of the students used the same formula format in both their examples, that is, they did not activate DoPV3.

3.1.3 Potential redesign by changing the type of example-generation task

We start by elaborating on how task 1 (type 1) could be converted into the second type of example-generation task (type 2). Figure 10.2 introduces a suggestion for redesign of task 1. Instead of asking for two distinct examples that fulfil certain conditions, students are encouraged to provide examples in response to prompts where various conditions are progressively added.

Below are some possible function properties (i) - (iv)

i. $f(x)$ is a rational function
ii. The degree of both the numerator and the denominator of $f(x)$ is 2
iii. $f(x)$ has two vertical asymptotes, $x = a$ and $x = b$
iv. $f(x)$ has a horizontal asymptote, $x = c$

a. Give an example of a function $f(x)$ satisfying (i) and (ii).
b. Give an example of a function $f(x)$ satisfying (i), (ii) and (iii).
c. Give an example of a function $f(x)$ satisfying the properties (i) - (iv).
d. Give an example of a function $f(x)$ satisfying all properties, except for (ii).
e. Give an example of a function $f(x)$ satisfying all properties, except for (i) and (ii).

Check in GeoGebra if your suggested functions really have the given asymptotes.

FIGURE 10.2 Suggested redesign of task 1 by changing the task type.

We assume that most students might provide rational functions in the form of a single quotient when responding to prompt (a). To respond to prompt (b), students need to adjust the denominator while likely maintaining the same function formula format. For prompt (c), where the goal is to meet all constraints, we assume that most students will adjust the coefficient of the x^2 term in the numerator. Due to constraints (i) and (ii), the type of function and degrees of the numerator and the denominator are fixed in the first three prompts. In prompt (d), the goal is to encourage students to activate the DofPV related to the numerator and denominator degrees (DofPV2 in Table 10.2). Finally, the intention behind prompt (e) is to prompt students to enrich their example spaces by creating non-typical examples, activating a DofPV not utilised by any of the students in the previous study (DofPV1 in Table 10.2). Compared to the original version of this task (see Figure 10.1), we claim that this revised version encourages students to enrich their example spaces. However, regarding the formula format (DofPV3 in Table 10.2), we predict that almost all students will use the same format (a single quotient) throughout prompts (a) to (d), thus not activating this DofPV.

3.1.4 Potential redesign by using adapted feedback

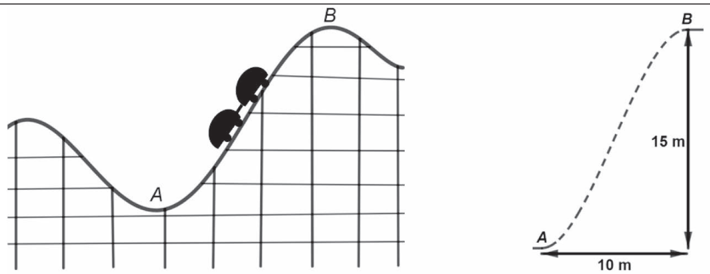
Next, we will discuss how the idea of ‘interactive example-generation tasks’ (Kinnear & Kontorovich, 2024) could be used in the redesign of the task. Since we assert that the revised version (in Figure 10.2) offers the potential for a richer example space, we will elaborate on this version of task 1. We believe it is instructive for students to recognise that different ways of thinking can result in various formula format. Hence, we propose asking for more than one example in prompt (c) (in Figure 10.2). However, requesting a second example might not encourage many students to activate DofPV3. This issue could be addressed with adapted feedback. For instance, if a student provides two valid examples written as single quotients (as we anticipate), the adapted feedback could be: 'Great, the answers are correct. Now, provide one more example with the horizontal asymptote as a single constant term.' To do this, it must be possible to use the CAA system to identify and compare the formats of the functions provided by the student. In STACK, this can be done using a formative potential response tree.

With these revisions of task 1, we claim that there is a good chance that many students will activate all DofPV in Table 10.2. They will probably activate DofPV4 when providing their second example for prompt (c), DofPV3 after receiving the adapted feedback for this response, and finally, DofPV2 and DofPV1 when responding to prompts (d) and (e), respectively.

3.2 Task 2

3.2.1 The original task

This is a modelling task (see Figure 10.3) of type 1, where students must first translate a real-life situation into mathematical conditions. Then, they are required to find two different functions that satisfy these conditions. The task is assigned as a group task.



Part of a structure in a roller coaster need to be replaced. It is the part between points *A* and *B* (see figures above) that should be replaced. Your task is to determine two different function formulas (*f* and *g*) that can be used during construction so that there is a smooth transition in points *A* and *B*.

- Use the measurements in the sketch to the right.
- Use GeoGebra to check if your suggested expressions are suitable before submitting them as answers to the task.
- There are a number of clues that you can use. However, it is important that you only use them when needed. So, try first to solve the problem without using any clue. If this does not work, then use as few clues as possible.

Clue 1: Start by placing points A and B in a coordinate system with the point A at the origin.

Clue 2: Use the conditions: f(0)=0, f(10)=15, f'(0)=0 and f'(10)=0

Clue 3: Suitable function types are polynomials of degree three and sine function

FIGURE 10.3 Original version of task 2, including the clues that are hidden.

As in task 1, students are encouraged to use a DMS to verify their suggestions before submitting them to the CAA system. In this way, they are prompted to use the instant DMS feedback, provided as visual information, for verification. Even though several types of functions could be utilised, students were expected to provide one trigonometric function and one polynomial function of degree 3. Since there are only one trigonometric and one cubic function that meet all conditions, the type of function is the only DofPV that students were supposed to activate.

3.2.2 The collective and situated example spaces observed

In the 2020 cohort, 98 groups of students (2–3 students in each group) responded to the task. As many as 59 groups (60%) provided one sine function and one cosine function, that is, various ways of expressing the same function. Accordingly, they did not activate the anticipated DofPV. In addition, only 14 (14%) of the groups provided the intended response, that is, one trigonometric function and one polynomial function of degree 3.

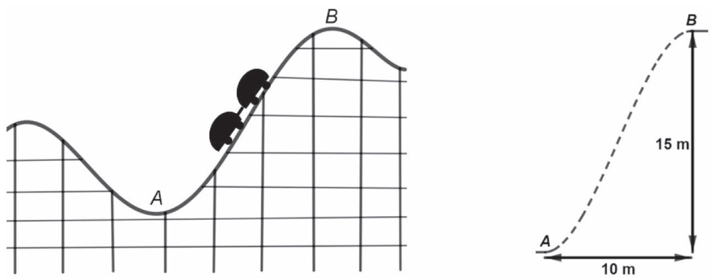
In the 2022 cohort, the task was divided into two separate tasks, one explicitly asking for a polynomial function and one requesting a trigonometric function. If students provided an incorrect answer, they received feedback from the CAA system in the form of clues before being given a second chance to answer the questions. Almost all students (95%) provided a correct polynomial function on their first or second try. Similar results also applied to the trigonometric task, where 94% of the students provided a correct answer. The division of the original task into two separate tasks prompted students to generate two types of function, that is, to activate the intended DofPV. However, the specification of function types reduces the openness of the task, thereby removing an important modelling aspect.

3.2.3 Potential redesign by changing the type of example-generation task

The original version of task 2 is an example-generation task of type 1. As for task 1, we start by elaborating on how the task could be converted into the second type of example-generation task. In the suggested redesign of task 2 (see Figure 10.4), all constraints are related to the type of mathematical function.

Compared to the latest version of this task, the main difference is in prompt (c). This prompt encourages students to activate an additional DofPV related to polynomial degree. To further enrich students' example spaces, we decided to ask for two examples in prompt (c). For the first example, we expect students to choose a polynomial of degree 4. For instance, they could start by assuming that $f'(x) = ax^2(x-10)$ or $f'(x) = ax(x-10)(x+10)$, which will generate the graphs in Figures 10.5a and 10.5b, respectively.

For the second example, students have the option to either stick with degree 4 polynomials or change the polynomial degree. If they opt for a degree



Part of a structure in a roller coaster need to be replaced. It is the part between points *A* and *B* (see figures above) that should be replaced. Your task is to determine a function $f(x)$ that can be used during construction so that there is a smooth transition in points *A* and *B*. Use the measurements in the sketch to the right.

Below are some possible function properties (i) - (iii)

i. $f(x)$ is a trigonometric function
ii. $f(x)$ is a cubic function
iii. $f(x)$ is a polynomial function

a. Give an example of a function $f(x)$ satisfying (i).
b. Give an example of a function $f(x)$ satisfying (ii).
c. Give two examples of functions satisfying (iii), but not (ii).

Use GeoGebra to check if your suggested expressions are suitable before submitting them as answers to the task.

FIGURE 10.4 Suggested redesign of task 2 by changing the task type.

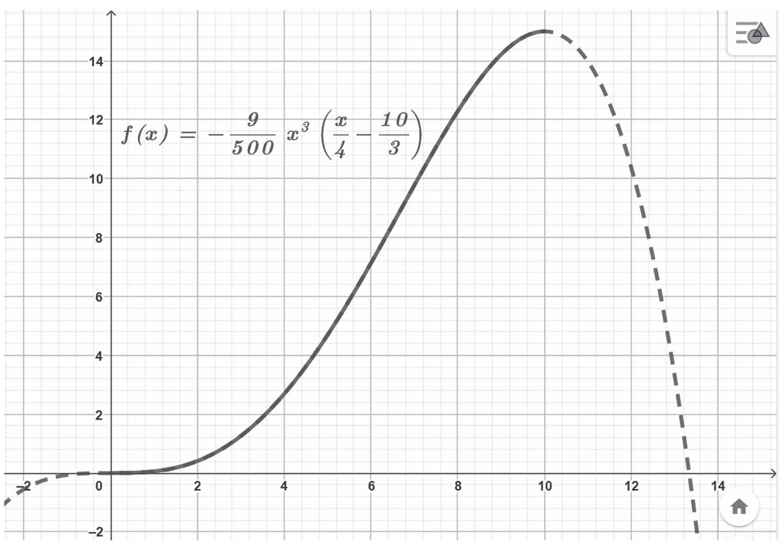


FIGURE 10.5A Response emanating from $f'(x) = ax^2(x - 10)$.

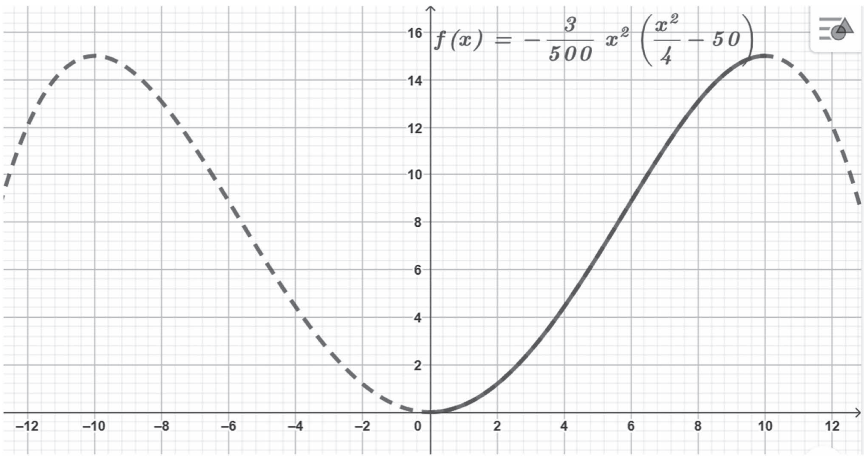


FIGURE 10.5B Response emanating from $f'(x) = ax(x-10)(x+10)$.

4 polynomial in their second example as well, they could start by choosing a proper value of the third zero b in $f'(x) = ax(x-10)(x-b)$, and then adjust the value of a to meet the conditions. On the other hand, students who assume that $f'(x) = ax^2(x-10)$ to create their first example might realise that the second example could be achieved by assuming that $f'(x) = ax^3(x-10)$, which will generate the function shown in Figure 10.6. Thus, in this manner,

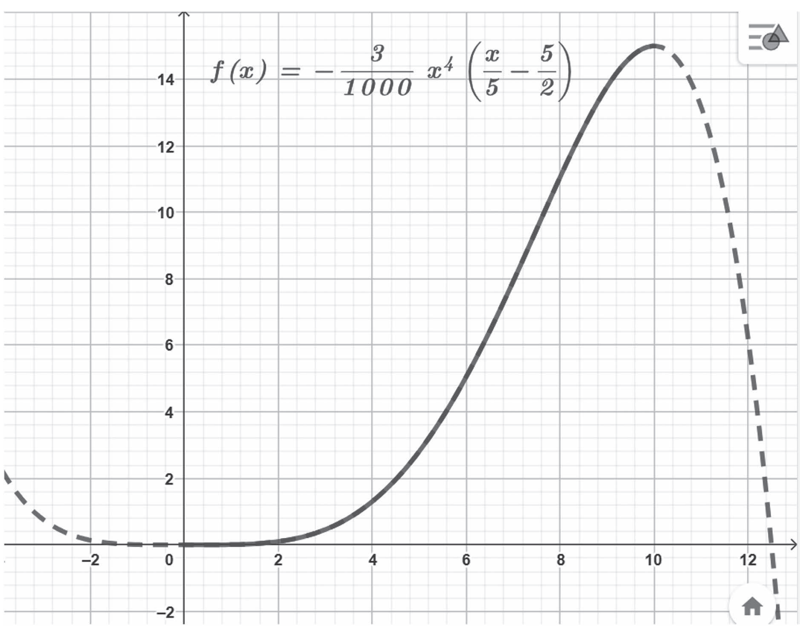


FIGURE 10.6 Response emanating from $f'(x) = ax^3(x-10)$.

these students utilise the polynomial degree as a DofPV. We claim that prompt (c) renders typical responses invalid and thereby encourages students to activate new DofPV, in this case the polynomial degree, and perhaps also the choice of a third zero.

3.2.4 Potential redesign by using adapted feedback

The revision for the 2022 cohort led to a high proportion of students providing both a correct cubic function and a correct trigonometric function. However, the revision reduced the openness by specifying the types of functions, consequently eliminating a crucial aspect of modelling. This drawback remains in the proposed redesign of the task in Figure 10.4. To address this issue, one approach could be to first ask for one example (without specifying the type) and then provide feedback tailored to the type of function (trigonometric or polynomial) provided. For instance, if the first example is a correct trigonometric function, the feedback could be 'Great, the answer is correct. Now, provide one more example that is not a trigonometric function'. This would address the problem observed in the 2020 cohort, where 60% of the students responded with one sine function and one cosine function. Likewise, if the first example is a correct polynomial function, the feedback could request one more example that is not a polynomial function.

After giving two examples, we assume that most students have provided one trigonometric function and one cubic function, which correspond to prompts (a) and (b) in Figure 10.4. As in the revised version in Figure 10.4, students' example spaces could be enriched by asking for two more polynomial functions of higher degree than 3. To further enrich students' example spaces, adapted feedback could be used. For students responding with two valid polynomial functions of degree 4, the feedback could be 'Great, the answers are correct. Now, provide one more example that is of a higher degree'. For students responding with degree 4 and degree 5 polynomials, the feedback could request one more example of a degree 4 polynomial. To give this kind of adapted feedback, it must be possible to use the CAA system to detect and compare the degrees of the provided polynomials.

3.3 Task 3

This task is adopted from Sangwin (2003) and involves a series of prompts introducing additional constraints progressively (type 2). The task is assigned individually, with the numerical values of the parameters (a) and (b) randomised for each student (see Figure 10.7).

The main key idea addressed in this task is the factor theorem, which involves comprehending how the zeros and factors of polynomials are interconnected.

Below are some possible properties (i) - (iv) of a polynomial $p(x)$.

i. $p(x)$ is a polynomial of degree three, i.e. $p(x)$ is a cubic function
ii. $p(a) = 0$ (iii) $p(b) = 0$ (iv) $p(0) = ab$

a. Give an example of a polynomial $p(x)$ satisfying (i).
b. Give an example of a polynomial $p(x)$ satisfying (i) and (ii).
c. Give an example of a polynomial $p(x)$ satisfying (i), (ii) and (iii).
d. Give an example of a polynomial $p(x)$ satisfying all the properties (i) - (iv).
e. Give an example of a polynomial $p(x)$ satisfying (ii), (iii) and (iv), but not (i).

Consider the possibility of testing your answers in GeoGebra.

FIGURE 10.7 Original version of task 3.

By adding constraints in terms of specific zeros for a polynomial function (as seen in prompt (b) and prompt (c)), the goal is to encourage students to apply the factor theorem. Moreover, prompt (d) introduces a further constraint by specifying the *y*-intercept, aiming to promote the use of vertical scaling – another key idea addressed in the task. Prompt (e) introduces a new DofPV in terms of polynomial degree.

3.3.1 The collective and situated example spaces observed

The main findings in the 2021 cohort, as reported in Fahlgren and Brunström (2022), show that the additional zero in prompt (c) led to a significant increase in students applying the factor theorem, that is, the main key idea. In this regard, the task worked properly. However, concerning prompt (d), the findings indicate that most students managed to respond without using vertical scaling (the other key idea), since their response to prompt (c), $p(x) = (x+1)(x-a)(x-b)$, also satisfied prompt (d). Concerning the last prompt, that is, providing an example of a polynomial function that fulfils all the given conditions except for being of degree 3, most students (80%) responded with the (only correct) second-degree polynomial, that is, $p(x) = (x-a)(x-b)$.

In the 2022 cohort, as reported in Fahlgren et al. (2024), the task was revised in two ways:

- To encourage students to use vertical scaling in prompt (d), we revised constraint (iv) into $p(0) = -2ab$.
- To enrich students' example spaces, we decided to ask for two examples in prompt (e). Since only one second-degree polynomial satisfies the given conditions, asking for another example will encourage students to explore polynomials with a degree of 4 or higher.

The findings indicate that the revision of constraint (iv) did not have the intended impact on student responses to prompt (d), since most students responded with $p(x) = (x-2)(x-a)(x-b)$, in factored or standard form. The revision of prompt (e) was effective in enriching the students' example spaces. Most students (59%) provided one polynomial of degree 2 and one of degree 4. Accordingly, these students activated the polynomial degree as one DofPV. There were also students (14%) who utilised this DofPV by providing polynomials of degrees 4 and 5. In total, 62% of the student responses to prompt (e) included the only correct quadratic polynomial $p(x) = -2(x-a)(x-b)$, either in standard or factored form. This result indicates that prompt (e) was effective also in promoting the use of vertical scaling.

3.3.2 Potential redesign by changing the type of example-generation task

In the original version of task 3, constraints are added progressively (type 2). In Figure 10.8, a revised version of task 3 into a task of type 1 is introduced. Similar to the original version of task 1 (Figure 10.1), students are required to provide two examples that meet all the given constraints.

Concerning the key ideas addressed by the task, no significant deviation is anticipated in the students' use of the factor theorem. However, the requirement for two examples may encourage students to employ vertical scaling. For instance, students who provide $f(x) = (x-2)(x-a)(x-b)$ as their first example, as observed in the 2022 cohort, might use vertical scaling when creating their second example. One potential disadvantage of this redesign, however, is that there are no prior responses to scale, as there is only one prompt incorporating all constraints.

Unlike the previous versions of the task, students are not asked to provide a non-cubic function. Consequently, polynomial degree is not utilised as a DofPV in this revised version of the task. To address this, the task could be reformulated by requesting two different polynomial functions, rather than cubic functions. Since only one quadratic function fulfils the given conditions, students providing this function would need to alter the polynomial degree to generate their second example.

3.3.3 Potential redesign by using adapted feedback

Building on previous findings from the 2021 and 2022 cohorts, as well as the discussion of potential redesigns through changes to the first type of example-generation task (Figure 10.8), we now discuss possible redesign by

> Give examples of two different cubic functions, *f* and *g*, both of which have
>
> - two zeros, $x = a$ and $x = b$, as well as
> - *y* intercept at $y = -2ab$
>
> Consider the possibility of testing your answers in GeoGebra.

FIGURE 10.8 Suggested revision of task 3 by changing the task type.

a. Give an example of a cubic function with two zeros, x = a and x = b.
[*Adapted feedback, such as "Great the answer is correct. Please provide one more example expressed in factored form", is provided for students responding in standard form*]
[*Adapted feedback, such as "Great the answer is correct. Please provide one more example", is provided for students responding with* $p(x) = x(x-a)(x-b)$]

b. Give two examples of cubic functions with two zeros, x = a and x = b, and y intercept at $y = -2ab$
[*Adapted feedback, such as "Great the answers are correct. Please provide one more example", is provided for students responding with*
$p(x) = (x-2)(x-a)(x-b)$ *and* $p(x) = -(x+2)(x-a)(x-b)$]

c. Give two examples of non-cubic polynomial functions with two zeros, x = a and x = b, *and y intercept at y = -2ab*
[*Adapted feedback, such as "Great the answers are correct. Please provide a quadratic function meeting the conditions", is provided for students responding with two non-quadratic polynomials.*

Consider the possibility of testing your answers in GeoGebra.

FIGURE 10.9 Suggested redesign of task 3 by using adapted feedback.

using adapted feedback. Figure 10.9 presents suggestions for further revision of Task 3. In this version of the task, a combination of both types is employed alongside adapted feedback.

The feedback for students responding in standard form to prompt (a) is provided to encourage them to use the factor theorem, one of the key ideas addressed by the task. The other feedback suggestions are provided to encourage students to use the other key idea, vertical scaling. The reason for the feedback in prompt (a), for students responding with $p(x) = x(x-a)(x-b)$, is that this response cannot be scaled to get the required y-intercept in prompt (b). In prompt (b), the reason for the feedback is that both $p(x) = (x-2)(x-a)(x-b)$ and $p(x) = -(x+2)(x-a)(x-b)$, that is, the reflexion of $p(x) = (x+2)(x-a)(x-b)$ in the x-axis, meet the conditions without the need for vertical scaling. In prompt (c), the feedback promotes vertical scaling by prompting students to provide a quadratic function, that is, $p(x) = -2(x-a)(x-b)$.

4 Conclusion

This chapter highlights the complexity of designing example-generation tasks. Designing tasks that foster a rich and varied example space while addressing key ideas can be challenging. Considering this challenge, we have emphasised crucial aspects to reflect on when designing example-generation tasks. We have redesigned three different tasks in various ways to illustrate how the design choices might affect the potential for encouraging a rich example space and addressing key ideas. In summary, our discussion on the three tasks suggests that:

- The choice to progressively add constraints to a sequence of prompts could enrich students' example spaces, as illustrated in the redesigned versions of task 1 (Figure 10.2) and task 2 (Figure 10.4). This is consistent with Watson and Mason's argument for this type of example-generation task. They assert that additional constraints can invalidate typical examples, thereby challenging students' creativity (Watson & Mason, 2005). Note that, in both task 1 and task 2, it is the condition that certain constraints should not be fulfilled that makes typical examples invalid.
- Tasks consisting of a sequence of prompts, where constraints are progressively added, can be improved by asking for more than one example in specific prompts, as demonstrated in the 2022 cohort of task 3 and illustrated in the redesigned version of task 2 (Figure 10.4).
- Using interactive example-generation tasks (Kinnear & Kontorovich, 2024) can enrich students' example spaces by activating further DofPVs while maintaining the task's openness. This is illustrated in task 2, where modelling is a key idea. In this task, we propose using adapted feedback to encourage students to reflect on suitable function types, instead of specifying which types to use.
- Asking for two examples and then providing adapted feedback, in which the request for a third example is based on the first two examples, might promote students to activate one further DofPV. This kind of interactive example-generation task is illustrated in the discussion on how to use adapted feedback in all three tasks.
- Interactive example-generation tasks can also enhance the opportunity for key ideas to be addressed within the task, without requiring too much tailoring. This approach is demonstrated in the redesigned version of task 3 (Figure 10.9), where adapted feedback is provided for specific responses in all three prompts.

To conclude, it is important to clarify that we have not tried these suggestions for redesign of the three tasks. In that sense, this is a speculative chapter. However, we claim that the chapter emphasises important aspects to reflect on when designing example-generation tasks, both in terms of the task's structure and the use of adaptive feedback. The common core of these aspects concerns when and how different constraints should be introduced and how many examples to request. In all three tasks, the redesign using adapted feedback illustrates the usefulness of letting the CAA system detect important features of the provided examples and then give new prompts based on these features. In particular, we claim that the notion of DofPV is very helpful when considering how to utilise the potential of a CAA system to provide adapted feedback on example-generation tasks. Since our suggestions for redesigning the tasks are based on speculations about expected student behaviour, a natural continuation would be to investigate how various design choices affect students' opportunities to enrich their example spaces and to apply key ideas.

Notes

1 www.digitaled.com/mobius/.
2 www.geogebra.org/.

Reference list

Antonini, S., Presmeg, N., Mariotti, M. A., & Zaslavsky, O. (2011). On examples in mathematical thinking and learning. *ZDM*, *43*(2), 191–194. https://doi.org/10.1007/s11858-011-0334-5

Bills, L., Dreyfus, T., Mason, J., Tsamir, P., Watson, A., & Zaslavsky, O. (2006). Exemplification in mathematics education. In J. Novotná, H. Moraová, M. Krátká, & N. Stehlíková (Eds.), *Proceedings of the 30th conference of the international group for the psychology of mathematics education* (Vol. 1, pp. 126–154). PME.

Brunström, M., Fahlgren, M., Vinerean, M., & Wondmagegne, Y. (2022). Designing for a combined use of a dynamic mathematics software environment and a computer-aided assessment system. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferrettti (Eds.), *Proceedings of the twelfth congress of the european research society in mathematics education (CERME12)* (pp. 3764–3771). Free University of Bozen-Bolzano and ERME.

Fahlgren, M., Barana, A., Brunström, M., Conte, M. M., Roman, F., Sacchet, M., Vinerean, M., & Wondmagegne, Y. (2024). Example-generation tasks for computer-aided assessment in university mathematics education: Insights from a study conducted in two educational contexts. *International Journal of Research in Undergraduate Mathematics Education*, *10*(3), 732–753. https://doi.org/10.1007/s40753-024-00252-4

Fahlgren, M., & Brunström, M. (2022). Example-generating tasks in a computer-aided assessment system: Redesign based on student responses. In H. G. Weigand, A. Donevska-Todorova, E. Faggiano, P. Iannone, J. Medová, M. Tabach, & M. Turgut (Eds.), *MEDA3 mathematics education in the digital age 3. Proceedings of the 13th ERME topic conference (ETC13)* (pp. 141–144). Constantine the Philosopher University in Nitra and ERME.

Fahlgren, M., & Brunström, M. (2023). Designing example-generating tasks for a technology-rich mathematical environment. *International Journal of Mathematical Education in Science and Technology*, *56*(3), 515–531. https://doi.org/10.1080/0020739X.2023.2255188

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., Foster, C., Heck, A., Henderson, K., Hunt, T., Iannone, P., Kontorovich, I., Larson, N., Lowe, T., Meyer, J. C., O'Shea, A., Rowlett, P., Sikurajapathi, I., & Wong, T. (2022). A collaboratively-derived research agenda for e-assessment in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education*, 1–31. https://doi.org/10.1007/s40753-022-00189-6

Kinnear, G., & Kontorovich, I. (2024). Using e-assessment for interactive example-generation tasks. In P. Iannone, F. Moons, C. Drüke-Noe, E. Geraniou, F. Morselli, K. Klingbeil, M. Veldhuis, S. Olsher, C. Hankeln, & P. Gonscherowski (Eds.), *Proceedings of FAME 1– feedback & assessment in mathematics education (ETC 14)* (pp. 158–165). Utrecht University.

Marton, F., & Booth, S. (1997). *Learning and awareness*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203053690

Mason, J. (2011). Phenomenology of example construction. *ZDM*, *43*(2), 195–204. https://doi.org/10.1007/s11858-010-0297-y

Moreno-Armella, L., Hegedus, S. J., & Kaput, J. J. (2008). From static to dynamic mathematics: Historical and representational perspectives. *Educational Studies in Mathematics*, *68*(2), 99–111. https://doi.org/10.1007/s10649-008-9116-6

Olsson, J. (2018). The contribution of reasoning to the utilization of feedback from software when solving mathematical problems. *International Journal of Science and Mathematics Education*, *16*(4), 715–735. https://doi.org/10.1007/s10763-016-9795-x

Rønning, F. (2017). Influence of computer-aided assessment on ways of working with mathematics. *Teaching Mathematics and its Applications*, *36*(2), 94–107. https://doi.org/10.1093/teamat/hrx001

Sangwin, C. (2003). New opportunities for encouraging higher level mathematical learning by creative use of emerging computer aided assessment. *International Journal of Mathematical Education in Science and Technology*, *34*(6), 813–289. https://doi.org/10.1080/00207390310001595474

Sangwin, C. (2013). *Computer aided assessment of mathematics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

STACK Docs. (n.d.). *Formative potential response trees*. https://docs.stack-assessment.org/en/Authoring/Potential_response_trees/#formative-potential-response-trees

Watson, A., & Mason, J. (2005). *Mathematics as a constructive activity: Learners generating examples*. Routledge. https://doi.org/10.4324/9781410613714-11

Yerushalmy, M., Nagari-Haddif, G., & Olsher, S. (2017). Design of tasks for online assessment that supports understanding of students' conceptions. *ZDM*, *49*(5), 701–716. https://doi.org/https://doi.org/10.1007/s11858-017-0871-7

Zaslavsky, O., & Zodik, I. (2014). Example-generation as indicator and catalyst of mathematical and pedagogical understandings. In Y. Li, E. Silver, & S. Li (Eds.), *Transforming mathematics instruction* (pp. 525–546). Springer. https://doi.org/10.1007/978-3-319-04993-9_28

Zazkis, R., & Leikin, R. (2007). Generating examples: From pedagogical tool to a research tool. *For the Learning of Mathematics*, *27*(2), 15–21.

11

STUDENTS' INTERACTION WITH AND APPRECIATION OF AUTOMATED INFORMATIVE TUTORING FEEDBACK

Gerben van der Hoek, Bastiaan Heeren, Rogier Bos, Paul Drijvers, and Johan Jeuring

Acknowledgements

We are very sorry that co-author Bastiaan Heeren passed away before the publication of this chapter, and we are grateful for his ability to articulate our thoughts, sometimes even before we had them.

1 Introduction

Nowadays, computer-aided assessment in mathematics education is used more and more (Sangwin, 2015). For example, computer-aided assessment systems can aid the learning process by providing error-specific feedback. Designers of computer-aided assessment systems face many choices with respect to feedback delivery. Narciss (2012) defines a *feedback strategy* as the specification of the way feedback is delivered within a learning environment. She singles out a specific feedback strategy, namely, the informative tutoring feedback strategy (ITF strategy). In an ITF strategy, a student does not immediately receive information about the correct response but is offered the opportunity to retry a task, to apply feedback information, such as error-specific hints. As such, an ITF strategy specifies a form of guidance for a student in the environment, where the student is provided opportunities to correct errors rather than study a correct solution.

There is an ongoing debate about the role of guidance during learning processes. In one corner we have, for instance, Kirschner et al. (2006) advocating direct instruction and worked examples. In the opposing corner, we have, for instance, de Jong et al. (2023) arguing in favour of inquiry-based approaches. Our goal is to use insights from this debate to set up an ITF strategy which

DOI: 10.4324/9781003533764-11

provides error-specific feedback and suitable subtasks to work on next. We aim to let students solve tasks without worked examples but by constructing the solutions themselves. However, in some cases, students may have difficulty starting a task without direct instruction; therefore, the environment offers optional direct instruction in the form of a video. In later stages, students can view worked-out solutions to the tasks.

In this chapter, we expand on our previous work about student experiences while working in an online environment (Van der Hoek et al., 2024) using new data. We designed an ITF strategy to support students in our learning environment. We analysed how this strategy affects the interactions of students with the environment, and how students appreciate the support they receive. Student behaviour while working in automated tutoring systems has been studied quantitatively, for instance, by Köck and Paramythis (2011) and Vaessen et al. (2014). Such quantitative studies rely on indicators to model students' behaviour, such as time spent in different modes of the environment or transition probabilities between these modes. However, such models of learners' interactions are an approximation of a complex reality. This is why, in this chapter, we use a qualitative approach to describe what happens when students are working in the environment.

A calculation by a student who is working in our environment is diagnosed only by inputting a final answer; this has two advantages. Firstly, working with final answer evaluation does not require additional skills to use the interface, allowing students to practice mathematics authentically (Kieran & Drijvers, 2006; Russell et al., 2003). And secondly, with final answer evaluation, there is no need for input fields to assess intermediate steps that could provide a scaffolding effect for the task (Tacoma et al., 2020). To provide error-specific diagnoses in our environment, we use model backtracking (MBT) (Van der Hoek, 2022; Van der Hoek et al., 2025), a technique that diagnoses a final answer to identify errors a student has made throughout the computation. As such, it allows a student to work on a problem using pen and paper and input only their final answer to receive feedback on the steps in their calculation.

To study how students interact with our design and whether students appreciate it, we use a qualitative small-scale design study with post-intervention interviews. We discuss how senior general secondary students aged 15 to 17 years interact with the environment. Furthermore, through the post-task interviews we investigate students' appreciation of different forms of feedback.

2 Theoretical framework

As with any learning, online learning requires guidance. In an online learning environment, this guidance can be provided as automated feedback. In what follows we discuss several feedback types relevant to an ITF strategy before moving to the role of guidance. We provide arguments from both sides of the discussion on guidance to argue that a balanced feedback strategy might be a good approach to facilitate learning in an online environment.

A feedback strategy specifies the delivery of feedback through a learning medium. In this study, we apply an ITF strategy (informative tutoring feedback strategy) (Narciss, 2012). An ITF strategy does not immediately present the correct response but offers learners the opportunity to retry tasks to apply previously received feedback information. For such feedback strategies, Narciss distinguishes three important dimensions (see Table 11.1): (1) the nature and quality of a feedback strategy, (2) the situational conditions of the instructional context, and (3) the individual characteristics of the learner. The first dimension, defining the nature of the strategy, includes three facets: (a) functional aspects related to objectives, such as fostering self-guidance and sustaining persistence; (b) aspects related to the purpose of the feedback content, such as identifying discrepancies between the learners' performance and the expected performance; (c) and aspects related to the presentation of the feedback, such as the level of specificity. The second dimension, defining the instructional context, deals with the core features of the instructional approach. This includes, for instance, identifying learning obstacles and errors. In a broader scope, the instructional beliefs of the feedback designers are embedded in this dimension. The third dimension of learners' characteristics encompasses, for instance, prior knowledge and learning strategies. In what follows, we describe a theory that will allow us to operationalise each of Narciss's three dimensions. In the design section, we will revisit these dimensions to elaborate how each is instantiated in our feedback strategy design.

Shute (2008) and Narciss (2012) identify several types of feedback (see Table 11.2), such as verification, try-again, and elaborated feedback. *Verification* feedback provides learners with knowledge about a response's correctness, often referred to as knowledge of results (KR). *Try-again* feedback (TA) allows learners to provide a new response after some other type of feedback is provided. As for *elaborated* feedback, Shute distinguishes several variants, two of which are of interest here: topic-contingent feedback and feedback on bugs. *Topic-contingent feedback* is feedback about the topic that is being

TABLE 11.1 Dimensions of a feedback strategy according to Narciss.

Narciss's dimensions of a feedback strategy
1. Nature and quality of a feedback strategy
a. Functional aspects related to objectives
b. Aspects related to the purpose of the feedback content
c. Aspects related to the presentation of the feedback
2. Situational conditions of the instructional context
3. Individual characteristics of the learner

Source: Narciss (2012).

TABLE 11.2 Feedback types, abbreviations, and descriptions

Abbreviation	*Feedback type*	*Description*
KR	Knowledge of results	Correct or incorrect responses are marked accordingly.
TA	Try again	A student has the opportunity to try a task again after a failed attempt.
WE	Worked example	A student can study a worked example to a task.
ES	Error-specific	A student receives feedback that is specific for an error made.
DI	Direct instruction	The theory and procedures involved are explained by an expert. This could be a video recording.

studied, which could be a worked example (WE) of a task or direct instruction (DI); *feedback on bugs* is error-specific feedback (ES), which is based on a diagnosis of a learner's response. These five feedback types are incorporated into our environment.

When learners execute a task, they experience cognitive load on their working memory (Kirschner et al., 2006). This load can be reduced by using direct instruction or worked examples. For instance, Sweller et al. (1998) show that worked examples alleviated cognitive load for low-ability students. Studies on worked examples versus learning from solving problems in the previous century (Chi et al., 1989; Renkl, 1997) generally favour learning from worked examples. The cognitive load that task execution introduces can cloud the actual learning process. However, Chi et al. (1989) also reported that positive learning outcomes using worked examples strongly depend on a student's ability to self-explain the steps in the worked example. This shows that worked examples promote the learning process, but there is a danger in solely relying on them.

Learning is a form of self-development; it constitutes a positive change of behaviour and knowledge stored in the long-term memory. Meanwhile, opportunities for self-development can be diminished by too much guidance. Moreover, through minimal guidance, students can develop the ability to evaluate their solution processes (Goodman & Wood, 2004). However, a drawback of minimal guidance is that a learner may experience a feeling of uncertainty (Bordia et al., 2004; Fedor, 1991) that could diminish their motivation. Uncertainty occurs when a learner feels they do not have sufficient information about their performance relative to task demands. This uncertainty can lead to frustration and disengagement (Williams, 1997). Thus, there should be enough room to explore in a learning environment, but there should also be possibilities to resolve uncertainty.

In conclusion, when learners perform a task, they can experience cognitive load and uncertainty, which can be reduced by using direct instruction and worked examples. However, too much guidance can impede self-development. To develop self-guidance abilities, there should be enough room for exploration for students. Cognitive demands and room for self-development must be balanced in an ITF strategy.

3 Research questions

Considering the theoretical framework, we set up the ITF strategy in our environment as follows: The available implemented feedback types, KR (knowledge of results), TA (try-again feedback), and ES (error-specific feedback), should allow students to postpone the use of the implemented WE (worked examples). As such, we create opportunities for students to develop self-guidance skills through exploration, and to receive guidance in case of cognitive overload or uncertainty. DI (direct instruction) is available in the form of a video should a student not know how to start.

We investigate whether our ITF strategy contributes to fruitful student–environment interactions and whether students appreciate it. We define an *interaction* as an event where the environment provides feedback and a student responds to the feedback by a subsequent action.

To do so, we address two research questions:

1. How do students interact with the various feedback types in the informative tutoring feedback strategy?
2. Do students appreciate the use of error-specific feedback as opposed to worked-out solutions?

4 Methods

We use a qualitative small-scale design study with post-task interviews. In this section, we describe the design, the instrument, the participants and treatment, as well as the data collection, and the data analysis.

4.1 Design

The environment offers two topics: linear extrapolation and exponential extrapolation. These topics are part of the curriculum for Dutch senior general students in the social science stream. Students regularly make errors when solving tasks on these topics (Esteley et al., 2004; Van Dooren et al., 2005). Therefore, we designed an online environment for performing such tasks. The environment (in Dutch) is available through the following link:

https://ideastest.science.uu.nl/mbt-server/.

Both the linear and the exponential extrapolation topics have the same informative tutoring feedback strategy. We first elaborate on these common design features before elaborating on the topics' design. The common design is an operationalisation of Narciss's (2012) dimensions for feedback strategies; we will show how these dimensions are realised after elaborating on the common design.

When entering the environment, a student views an initial instructional video on how to work within the environment. Next, the student begins to work on the main task. After entering a result in the environment, it provides KR feedback, and when a more elaborate diagnosis of a student error is possible, it provides ES feedback. A student then has the option to try again to obtain TA feedback. The ES feedback in the main tasks has low specificity (i.e. verbally formulated suggestions), whereas the ES feedback in the subtasks has higher specificity (i.e. suggestions that may contain calculations). WE feedback, a worked-out solution, is available only after a student has selected a subtask. A student, however, is at liberty to immediately select a subtask and view the worked-out solution. Once a student returns to the main task, WE feedback will be available for the main task. In the main task, a student can receive DI (direct instruction) using an instructional video on the topic. This video may be viewed only once to prevent a student from using it as a worked example. The flowchart in Figure 11.1 provides an overview of the various student options.

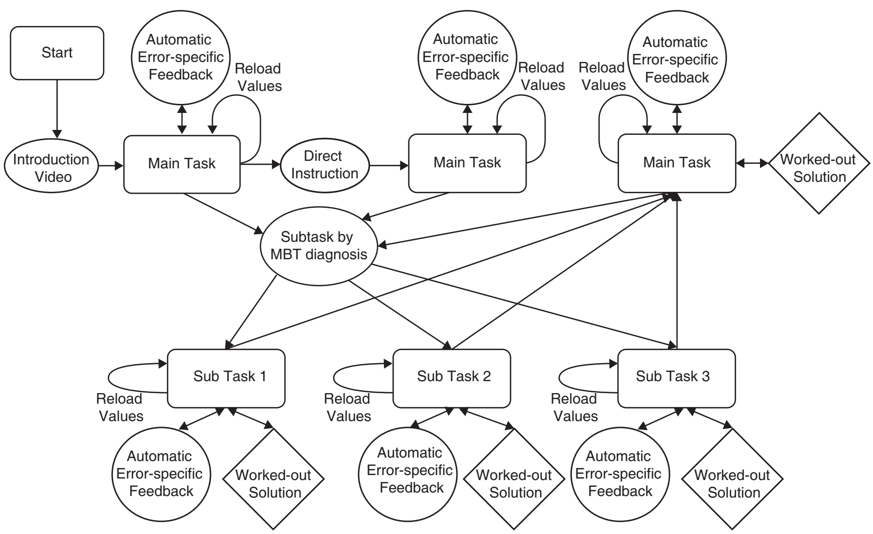


FIGURE 11.1 Flowchart of the ways in which students can navigate through the environment.

Now we describe how the three dimensions of Narciss are implemented; later, this implementation will be made more concrete when we discuss the learning environment. We start by elaborating on the three aspects of the first dimension about the nature of the feedback strategy:

a. Our main functional objectives are fostering self-guidance and maintaining engagement. To achieve self-guidance, various feedback types besides worked-out solutions are offered to allow room for exploration. Cognitive load and uncertainty are reduced by subtasks (ST), direct instruction (DI) (i.e. an instructional video on extrapolation), or worked-out solutions (WE).
b. The feedback content aims to enable students to identify discrepancies between their performance and the expected performance. To achieve this, the feedback either provides information on the specific student input (i.e. KR, TA, and ES) or has a high level of specificity (i.e. DI and WE).
c. The feedback presentation is such that the specificity increases at a student's request: The ES feedback has higher specificity in the subtasks, and a student has the option to view DI or WE.

Narciss's second dimension deals with the instructional context. In our case, we adopt opinions from both sides of the discussion between Kirschner et al. (2006) and de Jong et al. (2023) on guidance. This means that uncertainty and cognitive load are seen as the main learning obstacles and that worked examples should be avoided in the early stages to foster self-guidance. Moreover, allowing a student to choose a suitable form of help (i.e. ST, DI, WE) contributes to an ability to self-guide their learning process, although this is not further investigated in this study.

The third dimension encompasses learners' characteristics. In our case, the 15- to 17-year-old students have prior knowledge on the topics offered in the learning environment. However, it could be that this prior knowledge is insufficient to start a task; hence, the environment offers direct instruction in a video. Furthermore, we hypothesise that these adolescents tend to quickly resolve uncertainty through worked examples. Accordingly, we remove the option to view a worked example at the start of the task and only allow to view the direct instruction video once. In what follows we describe the implementation of the two topics: linear and exponential extrapolation.

4.1.1 Linear extrapolation tasks

For the linear extrapolation tasks, we set up a task format consisting of a main task and various subtasks (see Table 11.3). Students are first presented with the main task. When students input an erroneous solution for the main task, they receive feedback. After that, they can choose to correct their initial input or work on a subtask. In the latter case, the MBT system selects an appropriate subtask given the error.

TABLE 11.3 Tasks design structure for the case of linear extrapolation

<table>
<tr><th>Task</th><th>Example formulation</th><th>Additional description</th><th>Learning goal</th><th>Subtask selected in case of . . .</th></tr>
<tr><td>Main linear extrapolation task</td><td>Given the table:<table><tr><td>x</td><td>23</td><td>85</td><td>97</td></tr><tr><td>y</td><td>15</td><td>41</td><td>?</td></tr></table>Use linear extrapolation to compute the value of the question mark.</td><td>–</td><td>• Computing the slope (average rate of change)
• Using the slope to extrapolate</td><td></td></tr>
<tr><td>Subtask 1: simpler numbers</td><td>Given the table:<table><tr><td>x</td><td>54</td><td>55</td><td>93</td></tr><tr><td>y</td><td>64</td><td>57</td><td>?</td></tr></table>Use linear extrapolation to compute the value of the question mark.
To do so, first compute the change of y in a single step of x.</td><td>Task complexity is reduced compared to the main task since the x-coordinates of the given points are always consecutive and a way to start the computation is provided.</td><td>• Computing the change of y in this special case
• Using the change of y to extrapolate</td><td>• No input
• Undetectable error
• The student calculates: $? = 41 + (41 - 15) = 67$</td></tr>
<tr><td>Subtask 2: given slope</td><td>Given that the slope (rate of change) is equal to 8 for the following table:<table><tr><td>x</td><td>30</td><td>87</td></tr><tr><td>y</td><td>91</td><td>?</td></tr></table>Use linear extrapolation to compute the value of the question mark.</td><td>Task complexity is reduced relative to the main task by providing the slope.</td><td>• Using the slope to extrapolate</td><td>• Correct calculation of the slope (rounding errors allowed) and detectable error elsewhere</td></tr>
<tr><td>Subtask 3: computing slope</td><td>Given the table:<table><tr><td>x</td><td>35</td><td>62</td></tr><tr><td>y</td><td>47</td><td>68</td></tr></table>Compute the slope (the average rate of change).</td><td>Task complexity is reduced relative to the main task by requesting only the slope.</td><td>• Computing the slope</td><td>• Detectable incorrect calculation of the slope (rounding errors allowed)</td></tr>
</table>

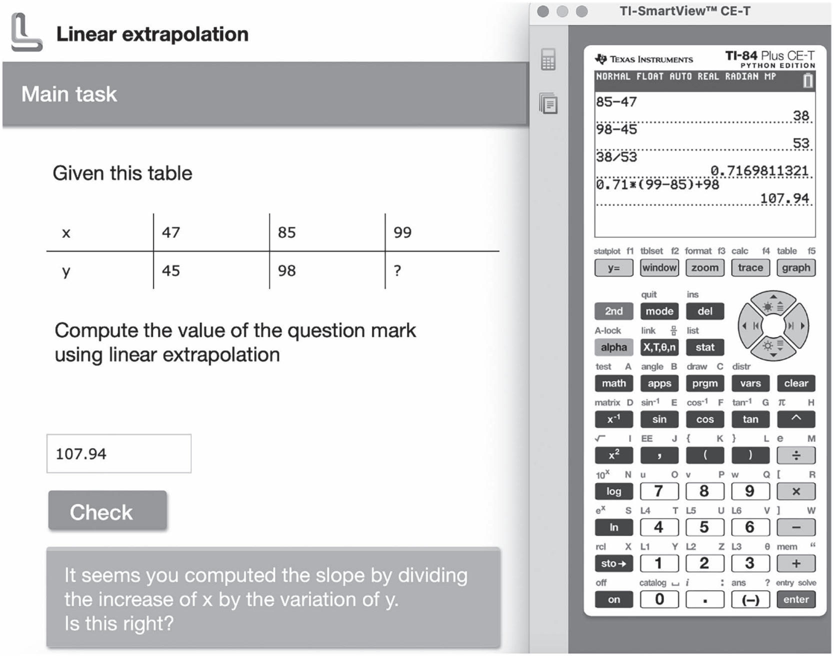


FIGURE 11.2 Example of error-specific feedback during a task.

The sets of numbers in each of the tables of the following tasks are so-called parameters. For each task, these parameters are randomly chosen from a set of 50 pre-calculated parameters that are tuned for final answer diagnosis accuracy using MBT techniques (Van der Hoek, 2022; Van der Hoek et al., 2025). This way, a student can retry a task with different starting values, while final answer diagnosis accuracy is maintained.

Figure 11.2 shows ES feedback a student receives when inversely computing the slope. To detect the various student errors, so-called buggy rules are implemented in the system. These buggy rules represent erroneous steps in a student's calculation (VanLehn & Brown, 1980). The rules are based on work by Van Dooren et al. (2005) on the unwarranted use of proportional models in missing value problems. Subsets of the buggy rules for the main task were used for the subtasks.

4.1.2 Exponential extrapolation tasks

Analogous to linear extrapolation, we formulate a task with subtasks for exponential extrapolation in Table 11.4.

For the exponential extrapolation task, the buggy rules are mainly based on work by Esteley et al. (2004) on using linear models in exponential situations.

TABLE 11.4 Task design structure for the case of exponential extrapolation

<table>
<thead>
<tr><th>Task</th><th>Example formulation</th><th>Additional description</th><th>Learning goal</th><th>Subtask selected in case of . . .</th></tr>
</thead>
<tbody>
<tr>
<td>Main exponential extrapolation task</td>
<td>Given the table:
<table><tr><td>x</td><td>77</td><td>80</td><td>85</td></tr><tr><td>y</td><td>58</td><td>55</td><td>?</td></tr></table>
Use exponential extrapolation to compute the value of the question mark.</td>
<td>–</td>
<td>• Computing the growth factor
• Using the growth factor to extrapolate</td>
<td>–</td>
</tr>
<tr>
<td>Subtask 1: simpler numbers</td>
<td>Given the table:
<table><tr><td>x</td><td>72</td><td>73</td><td>80</td></tr><tr><td>y</td><td>28</td><td>30</td><td>?</td></tr></table>
Use exponential extrapolation to compute the value of the question mark.
To do so, first compute the growth factor for a single step of x.</td>
<td>Task complexity is reduced relative to the main task since the x-coordinates of the given points are always consecutive and a way to start the computation is provided</td>
<td>• Computing the growth factor in this special case
• Using the growth factor to extrapolate</td>
<td>• No input
• Undetectable error
• The student calculates: $? = \frac{55}{58} \cdot 55 =$ 52.155</td>
</tr>
<tr>
<td>Subtask 2: given growth factor</td>
<td>Given that the growth factor is equal to 1.059 for the following table:
<table><tr><td>x</td><td>45</td><td>52</td></tr><tr><td>y</td><td>36</td><td>?</td></tr></table>
Use exponential extrapolation to compute the value of the question mark.</td>
<td>Task complexity is reduced relative to the main task by providing the growth factor</td>
<td>• Using the growth factor to extrapolate</td>
<td>• Correct calculation of the growth factor (rounding errors allowed) and detectable error elsewhere</td>
</tr>
<tr>
<td>Subtask 3: computing growth factor</td>
<td>Given the table:
<table><tr><td>x</td><td>28</td><td>40</td></tr><tr><td>y</td><td>72</td><td>34</td></tr></table>
Compute the growth factor for a single step of x.</td>
<td>Task complexity is reduced relative to the main task by requesting only the growth factor</td>
<td>• Computing the growth factor</td>
<td>• Detectable incorrect calculation of the growth factor (rounding errors allowed)</td>
</tr>
</tbody>
</table>

This environment was used to gather data on our research questions; in the next sections, we explain our data analysis setup.

4.2 Instrument

Below we present the post-task interview structure (see Table 11.5). Questions 1 through 7 are used to reflect on certain events during the session with the environment. They provide additional information for the first research question on the interaction with the environment. Questions 8 through 11 are used to gather information for future improvements of the environment. Finally, questions 12 and 13 are used to answer the second research question on the appreciation of an informative feedback strategy.

TABLE 11.5 Questions in the post-task interview

Interview
1. Did you understand the feedback you received?
2. Do you understand your error now?
3. Can you explain your error?
4. Can you explain why it was an error?
5. Was the subtask in line with the error you made?
6. Did the subtask help you understand the main task?
7. Did you understand the worked-out solution?
8. What do you think about the environment?
9. Are there cons to using this environment?
10. Are there benefits to using the environment?
11. What could be improved?
12. What do you prefer, to receive feedback on a single error or to view a worked-out solution?
13. Why?

4.3 Participants and treatment

For this qualitative research, 10 senior general secondary students from tenth grade and 15 from eleventh grade were recruited from six different classes in the school in the Netherlands where the first author is employed. Participation was based on availability and consent to partake.

The tenth grade students had received prior education on linear extrapolation as part of their standard curriculum. The eleventh-grade students had received prior education on linear and exponential extrapolation as part of their standard curriculum. However, the eleventh and tenth grade students did not receive instruction on these topics in the four weeks before the experiment.

Students were invited to complete the main task in the environment in a single session. The eleventh grade students could choose between linear and exponential extrapolation, or both when time allowed it. The tenth grade students worked on the linear extrapolation task only. Pen, paper, and an onscreen graphic calculator were available to the students. A researcher supported the students in case of confusion on how to operate the system, but not in case of confusion on the task. At times, the researcher reminded the students of the various options of the environment. After the end of the session with the environment, the researcher conducted a post-task interview with each student to determine the students' experiences with the environment.

4.4 Data collection

The data consist of screen-capture recordings along with the voices of the students and the researcher. One recording was not saved properly, but the researcher provided a written account instead. Therefore, we have data from 25 students navigating the environment. The duration of the navigation sessions ranged from 10 to 25 minutes and was sometimes restricted due to external factors, such as the start of the next class. Furthermore, we interviewed each student in 5- to 10-minute sessions, which were audio-recorded.

4.5 Data analyses

The data was analysed differently for the two research questions. For the research question on how the students interact with the ITF (informative tutoring feedback) strategy, units of analysis were identified. We define a *unit of analysis* or an *event sequence* as the student behaviour in between and including a starting state in Table 11.6 and a subsequent action in Table 11.7. The starting states and subsequent actions were chosen in this way because each navigating session can be covered by such units.

TABLE 11.6 Codes that signify the start of a unit of analysis

Code	*Current student state*
KR	The student receives KR feedback but no ES feedback.
ES	The student receives ES feedback.
mES	The student receives ES feedback as a result of a misdiagnosis.
WE	The student views the worked-out solution.
ST	The student returns to the main task from a subtask.
DI	The student views direct instruction.

TABLE 11.7 Codes that signify the end of a unit of analysis

Code	*Subsequent student action*
IM	The student improves the input relative to the last input for the same task.
nIM	The student does not improve the input relative to the last input for the same task.
WE	The student views the worked-out solution.
ST	The student chooses to work on a subtask.
DI	The student views direct instruction.
nCON	The student indicates not knowing how to continue.

If a student has no previous input, then any input is seen as an improvement. Additionally, improvement after viewing the worked example is measured through the first attempt with new starting values. Furthermore, when a student returns from a subtask, the last input in the main task is compared to the new input. A misdiagnosis occurs when the student input is not diagnosed properly.

If the allotted time expired during a unit of analysis, this unit was removed from the data. The transition frequencies from a student's state to a subsequent action are presented in Table 11.8. On 20% of the episodes resulting in either IM or nIM, a second rater independently agreed fully with the initial coding. Furthermore, the second rater independently reviewed the excerpts coded with nCON and fully agreed.

To investigate students' self-guidance ability, the students' calculations were examined for signs of testing the correctness of an intermediate result. That is, a student uses values in the task formulation to check whether an intermediate calculation step is correct. The test itself, however, need not be correct. Excerpts of the sessions were coded for this event (see Table 11.9). A second rater reviewed these excerpts and fully agreed with the initial coding. From the episodes, six episodes were selected exemplifying certain typical or atypical behaviours.

For the research question on students' appreciation of error-specific feedback as opposed to worked-out solutions, we analysed the post-task interviews. These opinions were prompted by questions 12 and 13 in the post-task interview. First, we coded the utterances on error-specific feedback and worked-out solutions using data-driven coding. We then grouped these initial codes and merged them in an axial coding process until an overview of themes emerged (see Table 11.11). A second rater coded 20% of the excerpts; Cohen's kappa[1] was $\kappa = .69$. After this initial coding, both raters discussed the differences, after which the raters agreed on this 20% of the data. Our sample size of 25 should be sufficient to achieve saturation of the themes in the interviews, because our population is fairly homogeneous, and the

TABLE 11.8 Frequencies of various event sequences

		Student's next action						
		Improvement	*Non-improvement*	*Worked example*	*Sub task*	*Direct instruction*	*Unable to continue*	*Total*
Current feedback	Only knowledge of results feedback	4	0	4	3	1	2	14
	Error-specific feedback	14	2	5	4	2	2	29
	Misdiagnosis	1	3	0	0	1	1	6
	Worked example	5	3	0	0	0	1	9
	Subtask	7	0	0	0	0	0	7
	Direct instruction	13	0	0	1	0	0	14
	Total	44	8	9	8	4	6	79

TABLE 11.9 Self-guidance frequency

Showed self-guidance	*Number of students*
No	20
Yes	5
Total	25

research object (i.e. the response to questions 12 and 13) is narrowly defined (Hennink & Kaiser, 2022).

5 Results

In this section, we present the results of our study. We first present the results on the interaction with the feedback strategy, followed by the results on students' appreciation of error-specific feedback as opposed to worked examples.

5.1 Results on interactions

The results relevant to the first research question on the interactions in the environment are presented in Table 11.8.

The improvement rate for direct instruction in this table is striking: 13 out of 14. This can be explained by the fact that direct instruction is often used at the start of a task, and no answer or a far-fetched answer is easy to improve upon. Another striking table entry is improvement through a subtask: 7 out of 7. This can be explained by the reduced cognitive load in the subtask in combination with the added presence of a worked-out solution allowing students to zoom in on their errors. Moreover, the subtask type is selected automatically based on the student error. The error-specific feedback shows improvement in roughly half the cases. In the other half of the cases, uncertainty is not removed sufficiently and students do not improve, seek additional feedback or even express an inability to continue.

The results of the investigation on self-guidance are presented in Table 11.7. It shows that 1 in 5 students displayed self-guidance by checking intermediate results.

From the units of analysis in Table 11.8 we collected six episodes and explained them through known theory; an overview is provided in Table 11.10. These episodes were selected because they illustrate certain typical and atypical interactions, and because during the post-task interview students made comments shedding further light on what happened during the episode. The episodes show that, generally, the feedback strategy offers guidance specific to an individual student's needs; still, in some cases, the environment does not provide the required guidance.

TABLE 11.10 Selected episodes and their theoretical interpretation

Episode	*Interaction*	*Theoretical interpretation*
Ian, learning through a subtask	Ian understood error-specific feedback on an error when he received it for the second time in the reduced setting of a subtask instead of in the main task. Upon returning to the main task, Ian solved it in a single attempt. In the post-task interview, Ian displayed insight into why his initial computation was erroneous.	Once the cognitive load (Sweller et al., 1998) was mitigated by the reduced complexity of the subtask, Ian could focus on his specific error.
Dani, not knowing how to start	Dani explicitly indicated she had no idea how to start; however, after receiving direct instruction, she could complete a large portion of the main task.	Direct instruction (Kirschner et al., 2006) reduced cognitive load (Sweller et al., 1998) and produced immediate results.
Leah, ES feedback	Leah received error-specific feedback messages, allowing her to complete the task without other forms of guidance.	Error-specific feedback (Shute, 2008) guided Leah towards a correct solution and produced direct results.
Yuna, self-check	Yuna tried to test whether an intermediate result was correct using values in the task, she explained in the post-task interview. As such, she exhibited self-guidance.	Exploration fostered a self-guiding ability (de Jong et al., 2023; Goodman & Wood, 2004).
Eve, memorising a worked example	Eve unsuccessfully tried to memorise a worked-out solution to a task. She repeatedly returned to the worked-out solution and talked about memorising it. However, she struggled to make sense of the task.	Eve lacked the proper self-explanation (Chi et al., 1989) skills needed to interpret a worked-out solution.
Robert, uncertain	Robert received knowledge of results feedback twice but could not correct his errors; eventually, he sighed. 'Now I have no idea what I did wrong.' He ceased his activity until prompted by the researcher.	The feedback was insufficient to alleviate Robert's uncertainty (Bordia et al., 2004; Fedor, 1991), causing him to become disengaged (Williams, 1997), until prompted by the researcher.

5.2 Results on student's appreciation of error-specific feedback

Concerning the results for the second research question on students' appreciation of error-specific feedback as opposed to worked examples, Table 11.11 shows the themes that resulted from the axial coding process. The themes *combination* and *self* have the highest frequency, showing that students appreciate a combination of first error-specific feedback and later worked-out solutions. This is in line with our ITF strategy setup. In the discussion section that follows, we will further analyse the contents of Table 11.11 in light of a known theory.

TABLE 11.11 Themes of students' opinions uttered during the post-task interview

Themes	*Description*	*Frequency*
Combination	First, specific feedback, then later the worked solution.	13
Self	If I view a worked-out solution, I don't have to do anything myself. *or* If I don't view a worked-out solution, I can still try it myself.	7
Overview	A worked-out solution provides an overview. *or* Specific feedback does not provide an overview.	6
How	Error-specific feedback does not always tell you how to proceed.	5
Understand	After studying a worked-out solution, I feel I understand how to solve the task, but when I have to do the task again, it turns out that I don't understand it.	5
Pinpoint	In a worked-out solution, an error is not pinpointed. *or* Specific feedback pinpoints an error.	4
Stop	After studying a worked-out solution, I don't want to continue because I think I understand.	4
Why	A worked-out solution does not always tell you why certain steps are made.	4
Total		48

6 Discussion

With respect to the first research question on the interaction with the informative feedback strategy (Narciss, 2012), we conclude from the results that the environment functions as desired. In Table 11.8 the total number of times a student either did not improve or indicated an inability to continue is 14. This leaves a total of 65 out of 79 successful transitions or improvements, where cognitive load (Sweller et al., 1998) and uncertainty (Bordia et al., 2004; Fedor, 1991) were such that students could seek additional guidance or improve their results. Moreover, Table 11.9 shows that, on average, 1 in 4 students showed self-guidance abilities.

In Table 11.10, Ian, Dani, Leah, and Yuna show how the informative feedback strategy guides the learning process. Furthermore, Table 11.8 and Table 11.9 show that these episodes can be seen as fairly typical since improvement occurred 44 out of 79 times. The episodes of Eve and Robert show that despite the efforts to remedy the lack of self-explanation skills (Chi et al., 1989) and prevent disengagement due to uncertainty (Bordia et al., 2004; Fedor, 1991; Williams, 1997), these effects still occur in the interaction with the environment. However, Table 11.8 shows that these events are somewhat atypical, since non-improvement occurred only 8 out of 79 times, and inability to continue 6 out of 79. Perhaps adding a prompt suggesting the use of direct instruction after several incorrect inputs could remedy this issue.

As for the second research question, on students' experiences with error-specific feedback and worked-out solutions, the results uncover two possible dangers of starting early on with worked-out solutions. Firstly, based on the *why* theme, with early use of worked-out solutions, students might not be ready to self-explain them (Chi et al., 1989). Secondly, from cases coded as *stop* and *understand*, we observe students can gain a false sense of certainty. This unjustified certainty causes them to be satisfied with the state of affairs and cease their learning activities, only to find out later they do not yet meet the requirements to solve similar tasks. This provides another argument to be careful with worked-out solutions during the learning process (de Jong et al., 2023).

The themes most mentioned are *combination* and *self*. They show that students appreciate a combination of first error-specific feedback and later worked-out solutions. This aligns with the main design principle of our feedback strategy, where room for exploration allows students to develop self-guidance abilities. This indicates that students appreciate a balanced, informative feedback strategy.

There are some limitations to the validity and generalisability of this study. The sample size is small, and the participants came from the same school in the Netherlands. This does not contribute to the generalisability of the findings. Furthermore, we do not measure any learning effects but, rather, analyse students' appreciation of the learning strategy. We do, however, offer explanations from known theory for the phenomena, which somewhat contributes to generalisability. Cultural influence and the inquiry-based book series used at the school may have influenced the results with respect to the second research question.

7 Conclusion

In conclusion, the results show that the ITF (informative tutoring feedback) strategy (Narciss, 2012) that balances guidance to accommodate students' individual needs is a strategy that students appreciate and that shows fruitful interactions. However, it remains a question of how this balanced ITF strategy would perform relative to a feedback strategy that, for instance, only provides worked-out examples and knowledge of results. Future research can shed light on such questions. Nonetheless, in this study, we showed that allowing room to explore (Goodman & Wood, 2004) while keeping cognitive load and uncertainty (Bordia et al., 2004; Fedor, 1991; Sweller et al., 1998; Williams, 1997) manageable promotes the learning process. The exploration fosters a self-guidance ability, while managing cognitive load and uncertainty allows students to improve themselves and remain engaged.

This shows that, for 15- to 17-year-old students, automated formative assessment benefits from a balanced informative tutoring feedback strategy for two reasons. Firstly, it allows room for exploration but can still provide the guidance needed for the individual student. Secondly, students appreciate receiving feedback through such a strategy. For developers of computer-aided formative assessment systems for this age group, these are important results to consider when deciding on an appropriate feedback strategy.

Acknowledgements

We are very sorry that co-author Bastiaan Heeren passed away before the publication of this chapter, and we are grateful for his ability to articulate our thoughts, sometimes even before we had them.

Note

1 Multiple codes could be assigned to a single excerpt; therefore, an additional category (theme) was added to the codebook, signifying a non-code. If the number of codes for an excerpt differed between raters, the smallest number of codes was supplemented with the non-code category, after which the number of codes for the excerpt is equal for both raters. Then, Cohen's kappa was calculated by viewing each code assignment as a single observation, allowing for multiple observations using the same excerpt. Here, each time the raters assigned the same category to an excerpt, it was viewed as a single agreement; the order in which the categories were assigned was immaterial.

Reference list

Bordia, P., Hobman, E., Jones, E., Gallois, C., & Callan, V. J. (2004). Uncertainty during organizational change: Types, consequences, and management strategies. *Journal of Business and Psychology*, *18*, 507–532. https://doi.org/10.1023/B:JOBU.0000028449.99127.f7

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

de Jong, T., Lazonder, A. W., Chinn, C. A., Fischer, F., Gobert, J., Hmelo-Silver, C. E., Koedinger, K. R., Krajcik, J. S., Kyza, E. A., Linn, M. C., Pedaste, M., Scheiter, K., & Zacharia, Z. C. (2023). Let's talk evidence – the case for combining inquiry-based and direct instruction. *Educational Research Review*, 100536. https://doi.org/10.1016/j.edurev.2023.100536

Esteley, C., Villarreal, M., & Alagia, H. (2004). *Extending linear models to non-linear contexts: An in-depth study about two university students' mathematical productions*. International Group for the Psychology of Mathematics Education.

Fedor, D. B. (1991). Recipient responses to performance feedback: A proposed model and its implications. *Research in Personnel and Human Resources Management*, *9*(73), 120.

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology*, *89*(5), 809. https://doi.org/10.1037/0021-9010.89.5.809

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, *292*, 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Kieran, C., & Drijvers, P. (2006). The co-emergence of machine techniques, paper-and-pencil techniques, and theoretical reflection: A study of CAS use in secondary school algebra. *International Journal of Computers for Mathematical Learning*, *11*(2), 205–251. https://doi.org/10.1007/s10758-006-0006-7

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction*, *21*, 51–97. https://doi.org/10.1007/s11257-010-9087-z

Narciss, S. (2012). Feedback strategies. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1289–1293). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_283

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*(1), 1–29. https://doi.org/10.1207/s15516709cog2101_1

Russell, M., Goldberg, A., & O'connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 279–293. https://doi.org/10.1080/0969594032000148145

Sangwin, C. (2015). Computer aided assessment of mathematics using STACK. In *Selected regular lectures from the 12th international congress on mathematical education* (pp. 695–713). https://doi.org/10.1007/978-3-319-17187-6_39

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296. https://doi.org/10.1023/A:1022193728205

Tacoma, S., Heeren, B., Jeuring, J., & Drijvers, P. (2020). Intelligent feedback on hypothesis testing. *International Journal of Artificial Intelligence in Education*, *30*(4), 616–636. https://doi.org/10.1007/s40593-020-00218-y

Vaessen, B. E., Prins, F. J., & Jeuring, J. (2014). University students' achievement goals and help-seeking strategies in an intelligent tutoring system. *Computers & Education*, *72*, 196–208. https://doi.org/10.1016/j.compedu.2013.11.001

Van der Hoek, G. (2022). Evaluating digital student work through model backtracking. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Twelfth congress of the european society for research in mathematics education (CERME12)* (No. 26, pp. 2873–2880). Free University of Bozen-Bolzano and ERME.

Van der Hoek, G., Heeren, B., Bos, R., Drijvers, P., & Jeuring, J. (2024). Students' experiences with automated final answer diagnoses for mathematics tasks. In P. Iaonnone, F. Moons, C. Drüke-Noe, E. Geraniou, F. Morselli, K. Klingbeil, M. Veldhuis, S. Olsher, H. Corinna, & P. Gonscherowski (Eds.), *FAME 1 – feedback & assessment in mathematics education (ETC 14)* (pp. 293–301). Utrecht University and ERME.

Van der Hoek, G., Jeuring, J., & Bos, R. (2025). Buggy rule diagnosis for combined steps through final answer evaluation in stepwise tasks. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education* (pp. 347–360). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98420-4_25

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for overgeneralization. *Cognition and Instruction*, *23*(1), 57–86. https://doi.org/10.1207/s1532690xci2301_3

VanLehn, K., & Brown, J. S. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379–426.

Williams, S. E. (1997, March). Teachers' written comments and students' responses: A socially constructed interaction. *Proceedings of the annual meeting of the Conference on College Composition and Communication,* Phoenix, AZ. Retrieved September 29, 2025, from https://files.eric.ed.gov/fulltext/ED408589.pdf

12

USING COMPUTER ALGEBRA TO SUPPORT AUTOMATIC ASSESSMENT OF MATHEMATICS

Christopher Sangwin and Michael O. Oyengo

1 Introduction to mathematical online assessment

Computer-aided assessment started in the 1960s, almost as soon as computers became available to educators; however, the origins can be traced back even further to programmed learning (Watters, 2021). An early example of computer-aided assessment is described by Hollingsworth (1960), who, with punchcard computers, tested students' programming (in machine language) in a computing class using a 'grader' programme:

> We could not accommodate such numbers without the use of the grader. Even though the grader makes the teaching of programming to large numbers of students possible and economically feasible, a most serious question remains, how well did the students learn? After fifteen months, our experience leads us to believe that students learn programming not only as well, but probably better than they did under the method we did use – laboratory groups of four of five students.
>
> *(Hollingsworth, 1960)*

The three sentences in this quote each contain issues still relevant today: large class sizes, a focus on what students actually learn, and the potential for serious learning gains.

Ambitious work in the 1980s was linked to research in artificial intelligence and was often titled *intelligent tutoring systems* (Sleeman & Brown, 1982; Quigley, 1988). These systems were pre-internet, often pre-network, and typically designed to support learning in one specific area, such as arithmetic, solving equations, or symbolic integration (Sleeman & Brown, 1982). We note

DOI: 10.4324/9781003533764-12

that some colleagues, notably Papert (1980), reacted against computer-aided instruction, claiming 'the computer is programming the child' when they preferred to encourage the child to programme a computer. The apparent dichotomy between *skills and their acquisition* and *conceptual understanding* appears to be perennial in mathematics education. Both extremes are unhelpful; indeed, this is probably a false dichotomy since concepts are difficult to understand without procedural skills. Whether the student is directed to complete structured tasks (on paper or online) or encouraged to work more freely (such as with project work or when programming) adds further complexity.

Notable assessment systems pre-2000 include CALM (Beevers et al., 1991, 1995), Diagnosys (Appleby et al., 1997), Cognitive Tutors (Anderson, 1986), Mathwise (Harding & Quiney, 1996), Aplusix (Nicaud et al., 2004), and many others. A particularly interesting example is the MathXpert, and its precursor, Mathpert, designed by Michael Beeson at San Jose State University (Beeson, 1998). The specific contribution of each of these projects to progress in this area was discussed in more detail in Sangwin (2013).

What distinguishes online assessment systems since 2000 from earlier work is the following:

1. Desktop applications have given way to networked (internet) systems.
2. Assessment systems are linked into larger general 'learning management' systems, typically run by institutions, which take care of user authentication and tracking of participation.
3. Especially since 2007 (the launch of the first iPhone), students increasingly access materials via mobile devices, rather than exclusively with desktop computers in specific computer labs. This freedom raises a number of issues, both in terms of security and the nature of the students' interaction with the assessments, both cognitive engagement and technical interface.
4. Older systems were (necessarily) fixed, focusing on specific topics. Contemporary tools let individual teachers author their own materials over a wide range of topics or select materials from mature and reliable question banks. Textbooks often come with online assessment support.

These changes are significant.

Despite being a desktop computer–based learning package, Mathwise was an early example where question authoring was via an authoring tool, rather than making the question designer write software for each item. The project produced 48 modules for topics typically taught in years 1 and 2 of university STEM courses (see Harding and Quiney, 1996). OpenMark was developed at the UK Open University, starting in 1974. This project ultimately made the transition from desktop to internet and allowed question authors to write their own materials, as described by Jordan (2011) and Jordan et al. (2011). An early example of an internet-based system is WeBWorK (Gage et al., 2002), which started at the University of Rochester in 1995 and is still in use in 2024. The software was popular with North American universities, and one distinctive feature was the Open Problem Library (Kehoe, 2010).

Each successive system has built on both the experiences of its predecessors and the availability of reliable general libraries and technology infrastructure. For example, last century, just displaying a traditional algebraic expression on a computer screen was something of a technical achievement, whereas now we have reliable libraries such as MathJax. From an educational perspective, Sangwin (2013) suggested the following common themes had emerged from these endeavours:

- Students appreciate the immediate feedback.
- Assessment of free-form steps in students' working were, in 2013, still difficult to automate, or tutorial systems working step by step are confined to specific topic areas. This is largely still the case in 2024.
- Input syntax is regularly cited as problematic by students, and an important learning goal by staff.
- Mathematical sophistication of the software is not tied to popularity.
- Replication and reinvention are rife.

2 The STACK online assessment system

This chapter reports the use of STACK (www.stack-assessment.org), a highly sophisticated assessment system. The original educational priorities for the STACK project were (1) using a computer algebra system (CAS) to give mathematical integrity to the assessment process and (2) allowing teachers to author their own questions. Political priorities were to make the software itself open-source, removing licence fees as a potential barrier, and encouraging shared community development. This freedom to extend and improve the software has been key in STACK's success. STACK is embedded within the Moodle and ILIAS learning management systems (LMS) so that many of the implicit decisions on assessment management are inherited from the quiz systems of the parent LMS.

In common with most current online assessment systems, the STACK question type has the following key features:

- Structured random question variants can be generated.
- STACK has a unique multi-part question design with a number of *input types*, such as algebraic expressions, multiple choice, and scientific units.
- Students typically provide an answer in the form of a mathematical expression.

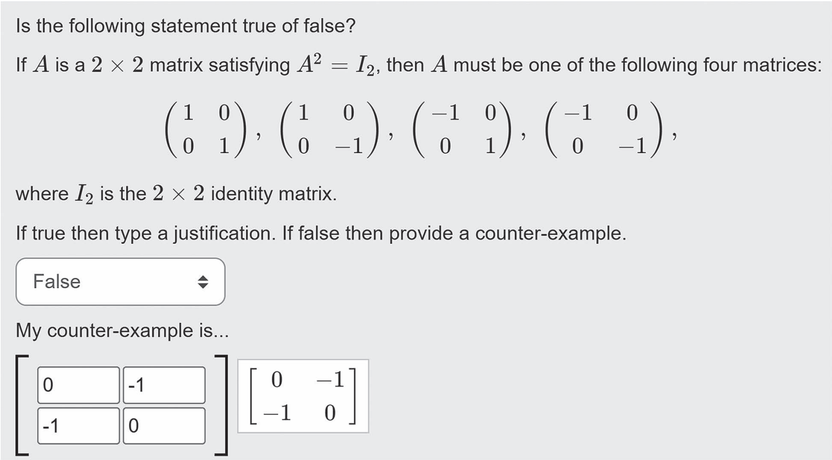


FIGURE 12.1 An example STACK question with students' answers.

- Students' input is carefully validated before assessment to prevent penalising a student for syntax errors.
- STACK establishes objective mathematical properties, for example, algebraic equivalence with a correct answer.
- Outcomes include numerical marks, formative feedback, and data on all attempts that are stored for later analysis.

A key strength is the ease with which a question author can include computer algebra calculations based on the student's answer in the feedback.

The example STACK question shown in Figure 12.1 illustrates some important features of STACK. Since this question is from an online examination, feedback on correctness would be deferred (and is not shown in the figure). This is a multi-part question, with some algebraic inputs and some multiple choice. In this example, students have been asked to choose true/false. The web pages adapt by revealing a matrix input for the student's counter-example. Had they responded 'true', they would have been given a text area to type in their justification. Here, immediate feedback is in the form of *validation*, echoing the student's answer and confirming there are no syntax errors.

Another example STACK question, shown in Figure 12.2, illustrates compatibility of STACK with other software, such as JSXGraph, to provide visual aid to students. This question is from a mastery quiz for formative assessment; feedback on correctness is personalised to the student's response, as shown in Figure 12.3. This question involves a random equation of a line from which students are to give three other lines, in the form of an algebraic input, so that they construct a square. The question tests the student's understanding of concepts of equations, parallel lines, perpendicular lines, intersection of lines, and distance between two points. These concepts are quite challenging to

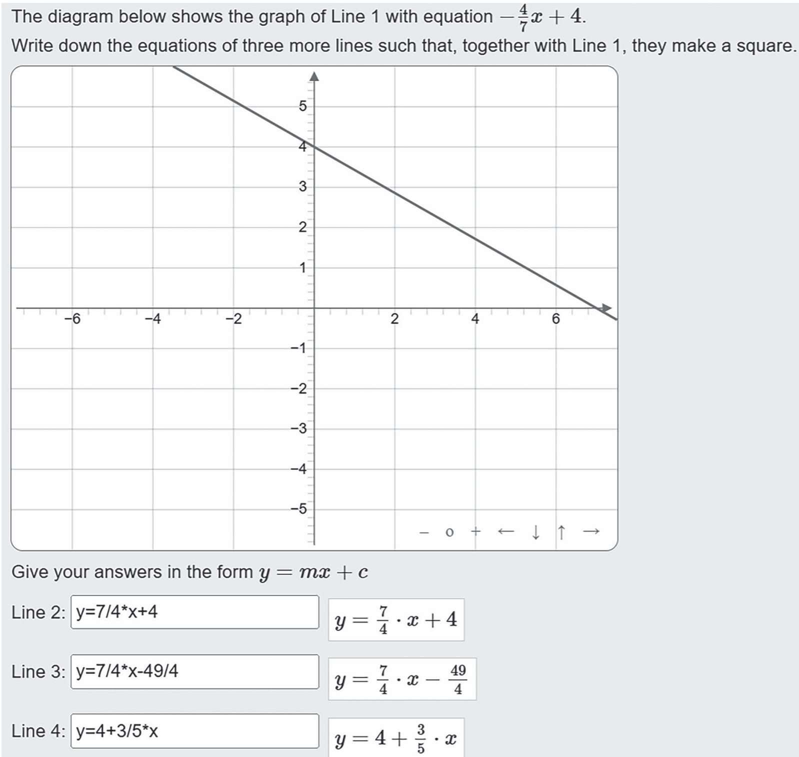


FIGURE 12.2 An example STACK question with algebraic answers.

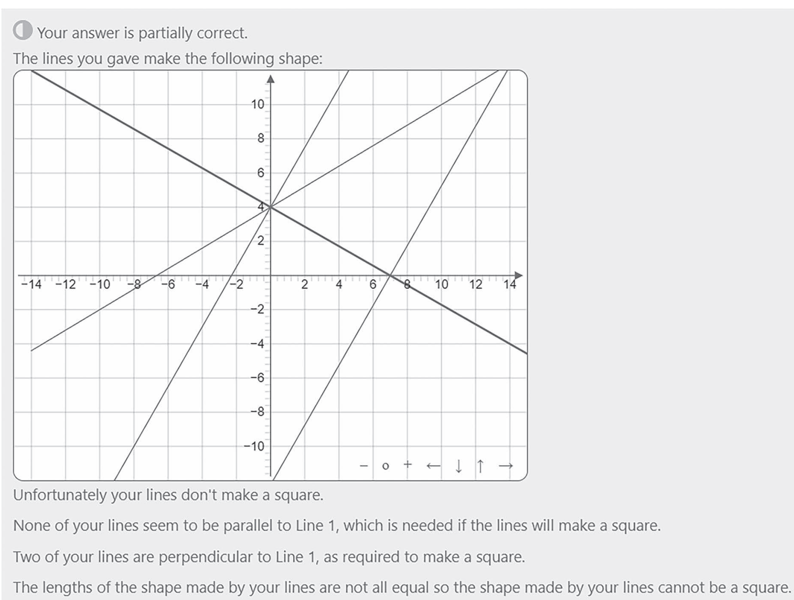


FIGURE 12.3 An example of detailed personalised feedback to a STACK question.

freshmen students, and the inclusion of a visual aid helps in the conceptualisation and understanding of the most important aspects of equations of a line. As in the previous example, immediate feedback is in the form of *validation*, echoing the student's answer and confirming there are no syntax errors. As can be seen in this example, even though the student's answer is 'incorrect', the system can still be designed to give partial credit. We find that personalised feedback, such as that shown in Figure 12.3, and randomisation of the question help motivate the student to reattempt it until they get it right. This promotes mastery through practice. STACK also provides teachers an opportunity to provide a full worked solution which reflects this randomisation (not shown).

A question type, such as STACK, is only one small part of the complex assessment process.

Individual questions are arranged into sequences, either a fixed 'quiz' or an adaptive test. These sequences contribute to a course or 'module' which, at university, is typically 8–11 weeks' duration. Then we have two further aspects of course design: (1) how knowledge is organised and how assessments relate to this organisation, and (2) examinations or other synoptic (rather than formative) assessment.

Questions need to be assembled into sequences, and students need managed access to these materials. There are two basic approaches to quiz management.

1. Questions are assembled into linear 'quizzes'. This replicates the traditional paper-based problem sheet, and while questions can be randomly selected from a question bank, or randomly generated from a template, the fundamental model is pre-determined by the teacher.
2. Questions are selected in an adaptive way, taking account of students' previous attempts. Adaptive learning typically requires some kind of user model and tracking of a student's previous attempts.

Adaptive learning is not new; indeed, the ideas can be traced directly back to ideas of programmed learning in the 1950s (Watters, 2021). Adaptive learning does not require computer support either; for example, adaptive materials have been developed in book format (Crowder & Martin, 1960). However, the clear lesson from previous adaptive learning projects is that developing adaptive materials is difficult, requires expertise on many levels, and is time-consuming. Typically, adaptive learning requires a substantial project, and few individual teachers have been able to create adaptive materials. Instead, individual teachers are more likely to be successful with the careful design of courses consisting of traditional quizzes.

The technical business of authoring an individual question really does affect all layers of this process. For example, if the only tool available is a multiple-choice question or other technically trivial assessment types, then the design of sequences of questions (including examinations) will be rather different from paper-based, open-ended assessments. Writing traditional examination questions includes many design decisions made to accommodate the assessment format and situation. Knowledge of how to use the tools available will influence the overall effectiveness of the course. The design of materials is the key here, rather than the technology used to deliver them. Indeed, Burkhardt (2012) argued for 'tests worth teaching to'. The next sections consider two contemporary examples where online assessment is at the heart of the overall assessment design.

3 Fundamentals of algebra and calculus

A notable example of effective assessment design is the course *Fundamentals of Algebra and Calculus* (FAC) (Kinnear et al., 2022). Developed by George Kinnear and Richard Gratwick at the University of Edinburgh, this course is delivered almost entirely online to year 1 undergraduate students. The original design goals echo the concerns of Hollingsworth (1960), namely, FAC needed to accommodate large class sizes, have a focus on what students actually learn, and be aimed at achieving serious learning gains. The course makes extensive use of the STACK computer-aided assessment, and its design was based on research from cognitive science and mathematics education research which had previously been found to be effective. In particular, the course is designed around the fundamental principle of *putting the textbook in the quiz*. By this we mean the online quiz is used to organise all course materials. All course materials take the form of quizzes, including text and video expository materials. Embedding exposition adjacent to interactive questions is designed to test students' understanding immediately. Further quizzes test overall learning, combining topics in summative assessments. This approach is discussed more fully in Sangwin and Kinnear (2022). In FAC, topics are interleaved (algebra/calculus/algebra . . .) because this distribution in time has been found to have a positive effect on how well the associated knowledge and skills are retained.

FAC is designed to improve students' understanding of calculus methods and supporting algebraic work to better prepare them for a traditional university calculus course. Unusually for a university course, FAC makes use of specifications grading. This is an approach in which individual assessments are judged as either pass or fail, with some opportunity for re-submission. The final grade for the whole course is then determined by the student's performance across these assessments, rather than by taking an average of numerical grades.

> Each weekly Unit Test was graded as Mastery (80%+), Distinction (95%+), or otherwise as a Fail. The final grade across the 10 weekly tests (accounting for 80% of the course grade) was then determined by the number of

> weekly units passed at which level. . . . In particular, to pass the course overall, Mastery level in at least 7 weekly units was required. Thus high expectations were set: even just a passing mark demanded a numerical outcome of at least 80% in 7 out of 10 weekly assessments.

In their evaluation of the course, Kinnear et al. (2022) found evidence of a positive impact on students as measured by pre-test and post-test learning gains. When provided with carefully designed and high-quality course materials, students rose to meet staff expectations.

This course design does not require online assessment, but online assessment provides automatic feedback to practice problems (which follow short video clips and text-based exposition) and allows students to sit the quizzes in a flexible time frame. Without online assessment, the course would be very difficult to deliver and repeated as needed. Such ideas trace their origin back to the mastery learning approaches of Bloom (1984).

4 The African context

In his World Bank report, Bethell (2016) gave a comprehensive analysis of the challenges facing mathematics education in sub-Saharan Africa. These challenges include inadequate teaching and learning materials, a shortage of qualified and trained teachers of mathematics, ineffective assessment systems that fail to provide mathematics teachers with the information they need to improve student achievement and instruction. Luneta (2022) notes that the COVID-19 pandemic only exacerbated these challenges in sub-Saharan Africa. One of the proposed solutions outlined by Bethell (2016) is the use of technology to support mathematics teachers and to harness the power of automatic assessment.

In Africa, there has been a steady increase in university enrolments (UNESCO, 2010), and yet government funding for public services in Africa has fallen. This is also happening elsewhere, including in Europe, but not so severely. Research conducted at Maseno University, Kenya, by Mukhanji et al. (2016) concluded that increased university enrolment led to 'unreasonably' large classes, inhibiting the interaction between students and lecturers. Indeed, most African public universities do not meet the minimum requirements for effective teaching, with lecture spaces, library, and laboratory resources being inadequate, unsuitable, and inaccessible by students and teaching staff. Mukhanji et al. (2016) concluded that there were no commensurate development and improvement of teaching and learning facilities to meet the rising enrolments, which has compromised the quality of teaching and learning.

Mathematics and other STEM subjects are most affected since effective learning relies on continuous interaction with content and immediate feedback to exercises (Hadijah et al., 2022). Many of these bottlenecks to mathematics education can be mitigated through the use of technology. It is therefore unsurprising that most African universities are exploring the use of technology in the teaching and assessment of mathematics, especially open-source learning and assessment materials that are accessible to low-resource environments.

For example, upon returning to Kenya in 2018 after his PhD studies in the USA, Michael Oyengo was faced with insurmountable challenges at Maseno University, where he teaches mathematics. He was assigned a linear algebra class of about 900 students to teach without additional support from teaching assistants. The teaching halls were too small, and the library was not equipped with sufficient resources to support the class. This was quite different from his experience in the USA, where he was one of the 16 teaching assistants supporting two professors to teach a similar course to a class of 1,000. As part of the teaching and assessment of the course at Maseno, he had to give two paper-based continuous assessment tests (CATs) and the end of the semester exam. The main challenge was giving meaningful formative feedback to the CATs in a timely manner to support learning. With the amount of marking required and without having additional support, giving individualised feedback and support to students in a timely manner was impossible. The best he could do was provide a grade to the CATs, and monitoring student learning and engagement was impossible.

The potential for the use of technology was clear, albeit he faced some serious initial challenges with the implementation, including the following:

1. *Technological.* Installation of the STACK plug-in on the Maseno LMS required technical expertise that was not available, and the university was not ready to meet the cost.
2. *Cultural.* Even though the department gave permission to pilot the use of technology, most lecturers were not ready to embrace technology and preferred the traditional paper exams.
3. *Access*. There was poor internet connectivity at the university, and students were forced to purchase internet data bundles.
4. *Devices.* Most students cannot afford laptops and tablets. Some did not have smartphones and had to share with their colleagues to access the homework. Table 1 shows how students accessed devices during the initial implementation of STACK at Maseno in 2019, from 443 responses (students were able to indicate multiple access methods).
5. *Authoring questions.* Lecturers did not have the skills or the time necessary to author STACK questions for the initial piloting phase.

To a large extent, these challenges are faced by all innovators, including those in better-resourced environments. For low-resource environments, the lack of resources forces innovation with online assessment since students will receive no feedback otherwise. For high-resource environments, the cultural, technological, and authoring challenges remain, but resources are often available to mitigate them.

Despite the initial challenges, STACK was successfully piloted at Maseno in 2019 for two courses: an introduction to linear algebra, and differential calculus. Some of the enablers for the initial pilot included the following:

1. Support by the University of Edinburgh through the use of their UK-based STACK-enabled servers.
2. IDEMS International (a not-for-profit community company based in the UK), which provided initial training to African lecturers on authoring and editing questions. They also provided direct support in authoring STACK questions and designing the assessments.
3. Students could access the online assessments through their smartphones, and those who did not have smartphones could borrow from their friends.
4. The university had a campus-wide wireless network that students could access from their devices. In some cases, students opted to purchase data from mobile phone network providers.
5. Departmental support was very important, especially in getting students to accept the mode of assessment.

For the initial implementation of STACK at Maseno, the two courses were divided into ten modules, with each module having two quizzes, a mastery quiz, and an end-of-module quiz. The mastery quiz was for practice and permitted an unlimited number of attempts, with the highest score being recorded for formative assessment. The end-of-module quiz could only be accessed once a student had a score of 70% on the mastery quiz, it could be done only once and had a time limit. Both quizzes contributed equally to the CATs for the two courses, and the mastery quiz could also be used for revision for the final exam. This pilot project was considered a success, with students recording an overall improvement in performance. Figure 12.4 gives an analysis of the final results with the quiz results. We note from the scatter plot

TABLE 12.1 How 443 students accessed the STACK website at Maseno University

Access method	*Frequency*
Use my own smart phone	417 (94.1%)
A friend's smart phone	48 (10.8%)
My own laptop	36 (8.1%)
A friend's laptop	21 (4.7%)
My own tablet	12 (2.7%)
A friend's tablet	1 (0.2%)
A computer in the mathematics laboratory	44 (9.9%)
Another desktop computer	20 (4.5%)

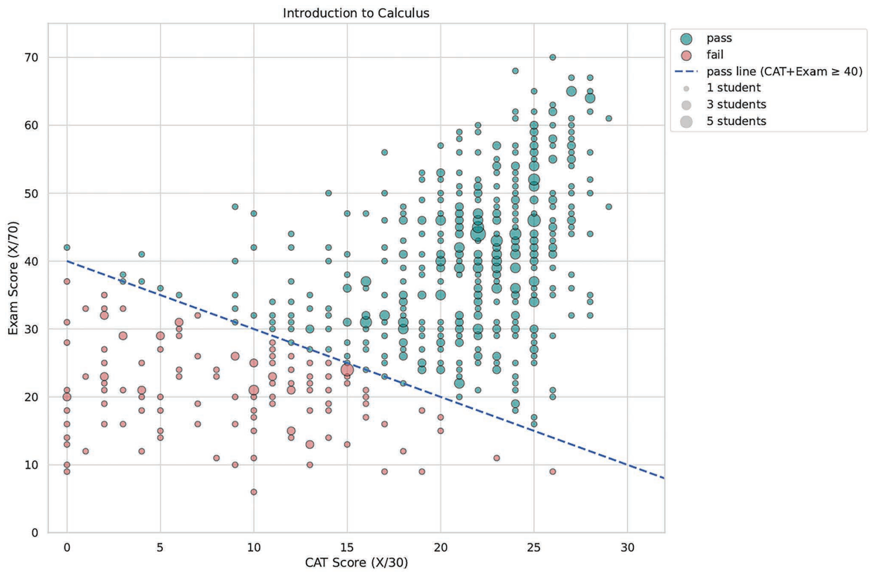


FIGURE 12.4 Student performance in the CATs (STACK quizzes) as compared to the final exam (paper exam) at Maseno University.

that most students who scored more than 40% on the CATs performed well on the final exam and passed the course. Conversely, most who performed poorly on the CATs (this was a very small percentage) did not do well on the final exam and failed the course. As would be expected, variation can be observed from the performance of students. In general, this analysis supports the hypothesis that the quizzes helped students in preparing for, and doing well in, their final exam.

For the first time at Maseno, students received immediate personalised feedback to their assignments and were encouraged to attempt the quizzes multiple times to build their mastery of concepts. There was an appreciation from students for this mode of assessment, with most of them grateful for the feedback that supported their learning process Oyengo et al. (2021). For staff involved, there was a sense of relief that hand-marking of paper CATs was replaced with online monitoring of student work, and their time could be directed instead to providing support to student questions on the online forums. Further, summary statistics could be used to easily identify problem areas for attention during class.

Just as with FAC in Edinburgh, the African experience echoes the original concerns of Hollingsworth (1960), namely, the need to accommodate large class sizes, a focus on what students actually learn, and the aim of achieving serious learning gains by students.

Key challenges experienced by university mathematics lecturers in the African context include the lack of time to learn how to (1) author (or edit) STACK questions, (2) improve basic programming knowledge, (3) write mathematics in LaTeX, and (4) manage teaching online courses. The initial success of STACK led directly to a workshop at Maseno in 2022, and then to the first pan-African STACK conference for undergraduate students in June 2023, at Masinde Muliro University of Science and Technology (MMUST), Kenya. The motivation for these meetings was to give a wider audience of African mathematics lecturers, educators, and researchers a local platform to present their work and engage one another. It became evident during these meetings of a need for a team of technical professional specialists trained in authoring STACK questions and supporting lecturers in implementing STACK at their universities. Later in 2022, IDEMS International, in collaboration with their partner company INNODEMS Kenya, created a STACK internship program where four recent university graduates were hired through a competitive application process and trained in authoring STACK questions. The trainers were IDEMS staff, supported locally by African lecturers who had the technical skills to verify the mathematical accuracy of the authored questions. This process was very costly, both in terms of time and money, for IDEMS and INNODEMS. The initial group of African STACK professionals consisted of the four interns and two further Maseno graduate students (who have subsequently gone on to their PhDs in Italy). The process of learning to author questions resulted in tremendous growth, both personally and professionally. They started as *consumers* of STACK quizzes throughout their four years as undergraduate students. Ultimately, the interns were heavily involved in organising the Africa STACK workshops and conferences, gaining valuable interpersonal skills and interacting with university lecturers. They now have enough experience to author quality STACK questions and support lecturers from various African universities with online course management.

The African experience mirrors that of colleagues in Europe, as has been acknowledged earlier, for example, Sangwin and Grove (2006). The technical process of automating online assessment is an additional technical challenge for, often busy, academic staff. The technical challenge is in addition to the challenges of designing effective assessments. Academic staff need autonomy; for example, they need to choose and adapt materials to their own courses. To fulfil this need, there are a number of choices. Staff can develop expertise themselves, they can collaborate with professional specialists, or they can *select* materials from pre-existing question banks.

5 The Open Question Bank (OQB) and shared digitised courses

In an environment where resources are scarce, it is imperative to combine efforts for the common good. For example, in Ethiopia, all 42 public universities offer a harmonised curriculum to their students. This has a significant practical implication: If a course is digitally developed for one university, all public universities can use that particular course with minimal changes. Many African public universities share similar curricula for their courses. For example, in Kenya, the founding public universities were the University of Nairobi, Kenyatta University, Jomo Kenyatta University of Agriculture and Technology, and Moi University. The other 26 public universities were established from the founding universities and inherited their curricula. Once a mathematics course is developed online with STACK assessment resources, for instance, at Maseno University, it can be shared with minimal changes across other universities in Kenya. This means that a standard can be set for the quality of a course during its preparation, and other universities that adopt the course can attain or improve upon this standard by enhancing the course.

Developing a course online involves the following multi-step collaborative process.

Step 1: The course lecturer prepares draft questions that need to be digitised in STACK.

Step 2: A 'question conceptualisation' process is initiated by the lecturer, who provides the question text, a comprehensive worked-out solution, and possible mistakes students might make.

Step 3: Members of the STACK professionals team (who may or may not be mathematicians) first check the OQB for similar questions to avoid duplication. Either they adapt a similar question or author a new question following the concept document.

Step 4: Questions are reviewed by the STACK professionals. This includes grammar, that the question works properly with randomisations, and that it aligns solutions with the concept document.

Step 5: A second review is done by the course lecturer to ensure mathematical accuracy. The lecturer verifies that the randomisations fit the question requirements and that different variants of the question have similar difficulty. The lecturer can accept the question or request further changes.

Step 6: The question is now ready for deployment in the course, for example, for use in a quiz on the university's learning management system (LMS).

Step 7: Data of use by students is collected and analysed to assess the question's usefulness, identify common student mistakes, improve feedback, etc. This data is used to improve the concept document and revise the question.

Emphasis is placed on developing assessment material for a particular course to be delivered in a blended format, combining face-to-face interaction with the course lecturer and online continuous assessment and monitoring of student learning.

From the Open Question Bank (OQB), questions can be used directly or adapted by the course lecturer to fit their context. The tension between writing materials from scratch, adapting existing materials, and finding and selecting from banks of existing questions occurs in all contexts, not just in Africa. Indeed, all users of online assessment, including STACK, face this problem. The WebWork community concentrated on developing open problem banks, rather than the mathematical sophistication, as a system priority. However, teachers often underestimate the time needed to review interactive materials, especially the feedback generated to different incorrect answers. Teachers also underestimate the value of *sequences of questions*, rather than individual questions, when sharing resources and setting up resource libraries. In Europe, early adopters have largely been fortunate enough to have had sufficient resources in terms of staff time and expertise to write their own materials from scratch. Or institutions have invested in creating new posts for technical experts to write such materials as part of a team with subject experts. Since we now have a wide variety of sophisticated and reliable assessment tools,

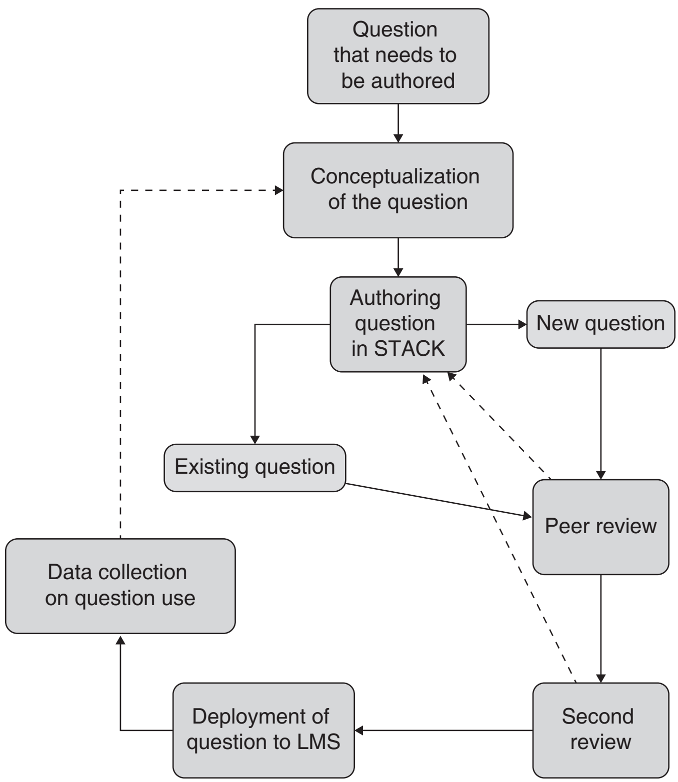


FIGURE 12.5 Question authoring process for the Open Question Bank (OQB).

attention is turning to the practical problem of sharing question banks using version control software (e.g. GIT) to manage and track contributions and changes.

6 Future use in the context of AI

Artificial intelligence has developed so that, in 2024, we have highly effective tools, including speech and text recognition, with the reliable conversion of images of handwriting to LaTeX mathematics, such as https://mathpix.com/. Grammar checkers and other writing assistants have advanced considerably. These specific AI tools are impressive, and useful. More recently, generative AI has entered the mainstream consciousness and started to be used in education. In the spring of 2023 Khan Academy incorporated an AI-driven student tutor and teacher's assistant they call Khanmigo. Automated chat assistants clearly have potential to help students but are not without some potential pitfalls:

> The tutors appear to work better if they present themselves to students as nonhuman tools to assist learning rather than as emulations of human tutors.
>
> *(Anderson et al., 1995)*

During the response to the COVID-19 pandemic, uptake of online assessment increased dramatically, and unfortunately, some colleagues had quite unrealistic expectations of what objective assessment is designed to achieve. For example, Sangwin and Bickerton (2023) discuss in detail why assessment of proof is surprisingly subjective, and so assessment of proof remains difficult to automate. AI *may* have a role for the assessment of longer answers, where assessment criteria are more subjective. However, we still have to decide what we will, and will not, accept and train the AI accordingly. Even then, individuals, and society as a whole, may not accept automated judgements by AI in high-stakes situations, such as formal examinations.

At the 2024 annual STACK user group meeting, Jesús Copado (SURLABS) demonstrated extensions of STACK to allow ChatGPT to help the question author automatically write complete worked solutions. These tools were capable of automatically generating complete STACK question XML code. We expect such tools to develop and considerably ease the difficulties of question authoring; however, ultimately the teacher remains responsible for the assessments in their course.

Adaptive tests have proved very difficult to write.

> Early CAI [computer-aided instruction] workers set themselves the task of producing teaching systems which could adapt to the needs of individual

> students. It is now generally agreed that this is a *very* difficult task, and one that will only be accomplished as a result of extensive research in both AI and Cognitive Science.
>
> *(Sleeman & Brown, 1982, Preface)*

Large classes and the ubiquity of current mathematics education mean we typically have extensive datasets in mathematics education, with which we may be able to train AI to support adaptive testing.

> The core commitment at every stage of the work and in all applications is that instruction should be designed with reference to a cognitive model of the competence that the student is being asked to learn. This means that the system possesses a computational model capable of solving the problems that are given to students in the ways students are expected to solve the problems.
>
> *(Anderson et al., 1995)*

This quote calls for mindfulness on the part of the teacher, much as we expect our students to engage with the assessments we set.

AI could prove useful in improving the quality of feedback and the quality of STACK questions. As we continue to gather large datasets of information from student interaction with STACK questions, we need to sift through the data and identify common student mistakes and misconceptions. This information will then inform the changes that need to be done on a particular question, whether it is the context, the feedback, or the language used within the question, so as to improve the learning experience.

There is considerable uncertainty about the role of AI in education. One dystopian future involves (nefarious) the use of AI by students to answer AI-generated questions. Clearly, this helps nobody: Students need to engage to learn, and society expects integrity from formal assessment systems. AI will not obviate the need to create challenges for students, or the need for students to engage meaningfully with these challenges. We predict that objective assessments automated by contemporary CAS-based systems will not entirely disappear, just as MCQ has not entirely disappeared. New AI tools will find their niche for effective use, and colleagues will select tools where they are most effective, rather than because they present themselves as the only option.

7 Conclusion

This chapter considered one important aspect of digital technology for assessment in mathematics education, namely, tools for automatic online assessment, as exemplified by the STACK system. The original concerns of Hollingsworth (1960) remain just as relevant today, namely, the need to accommodate large class sizes, a focus on what students actually learn, and the aim to achieve serious learning gains by students. What has changed during the last 60 years is ubiquitous internet connectivity and the ready availability to many students of hardware, especially mobile phones. High-quality online materials are now available for students, and this technology scales easily.

The challenges of, and opportunities offered by, automatic online assessment are shared internationally. In particular, the need for experienced mathematics teachers to learn how to use new tools, both at a technical level and effectively to support education, highlights an entirely new form of work which traditional teaching did not include. This work includes randomly generating questions, equivalent in educational purpose and within a coherent quiz experience for students. This work also includes establishing properties, deciding which detailed properties are relevant, and predicting what feedback to give in each outcome. As sophisticated and reliable tools become mainstream, the focus of attention is turning to question banks and how to develop and maintain reliable and effective libraries of questions. Even where reliable shared question banks exist and can be searched and re-used efficiently, teachers need to understand how to use and adapt materials for their own class. It is likely that the most efficient way to support teachers is to provide access to technical specialist professionals, rather than require every teacher to develop technical expertise. Science teachers have been supported by lab technicians for many years, and we suggest that mathematics teachers might also benefit from the support of online assessment technicians.

Contemporary online assessment is likely to remain a useful tool in education for the foreseeable future in parallel to more widespread AI use. Artificial intelligence is likely to make important contributions to education. Opportunities include automatic assessment of students' answers, of course, and in generating feedback to students. AI also has the potential to support teachers in writing questions (technical code generation), in selecting questions from large question banks, and in automatic translation between languages to help share materials internationally. Many of these opportunities will positively influence student outcomes; however, what continues to matter is what students actually learn.

Reference list

Anderson, J. R. (1986). *The architecture of cognition*. Harvard University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402n2

Appleby, J., Samuels, P. C., & Jones, T. T. (1997). DIAGNOSYS – a knowledge-based diagnostic test of basic mathematical skills. *Computers in Education*, *28*, 113–131.

Beeson, M. (1998). Design principles of mathpert: Software to support education in algebra and calculus. In N. Kajler (Ed.), *Computer-human interaction in symbolic computation* (pp. 89–115). Springer-Verlag. https://doi.org/10.1007/978-3-7091-6461-7

Beevers, C. E., Cherry, B. S. G., Foster, M. G., & McGuire, G. R. M. (1991). *Software tools for computer aided learning in mathematics*. Avebury Technical.

Beevers, C. E., McGuire, G. R., Stirling, G., & Wild, D. G. (1995). Mathematical ability assessed by computer. *Journal for Computers and Education*, *25*, 123–132.

Bethell, G. (2016). *Mathematics education in Sub-Saharan Africa: Status, challenges, and opportunities (tech. rep. ACS19117)*. The World Bank.

Bloom, B. S. (1984). The 2 sigma problem. *Educational Researcher*, *13*(6), 4–16.

Burkhardt, H. (2012). Engineering good math tests. *Education Week*, *32*(6), 23.

Crowder, N. A., & Martin, G. C. (1960). *Adventures in Algebra*. Doubleday.

Gage, M., Pizer, A., & Roth, V. (2002). WeBWorK: Generating, delivering, and checking math homework via the Internet. In *Proceedings of ICTM2 international congress for teaching of mathematics at the undergraduate level*. Hersonissos, Crete, Greece, July 1–6, 2002. University of Crete. Available at: http://www.math.uoc.gr/~ictm2/Proceedings/pap189.pdf

Hadijah, H., Isnarto, I., & Walid, W. (2022). The effect of immediate feedback on mathematics learning achievement. *Journal of Pijar MIPA*, *17*(6), 712–716. https://doi.org/10.29303/jpm.v17i6.4172

Harding, R., & Quiney, D. (1996). Mathwise and the UKMCC. *Active Learning*, *4*, 53–57.

Hollingsworth, J. (1960). Automatic graders for programming classes. *Communications of the ACM*, *3*(10), 528–529. https://doi.org/10.1145/367415.367422

Jordan, S. (2011). Using interactive computer-based assessment to support beginning distance learners of science. *Open Learning*, *26*(2), 147–164.

Jordan, S., Jordan, H., & Jordan, R. (2011). Same but different, but is it fair? An analysis of the use of variants of interactive computer-marked questions. In *Proceedings of international computer assisted assessment conference,* 5–6 July 2011, Southampton, UK.

Kehoe, E. (2010). AMS homework software survey. *Notices of the American Mathematical Society*, *57*(6), 753–757.

Kinnear, G., Wood, A. K., & Gratwick, R. (2022). Designing and evaluating an online course to support transition to university mathematics. *International Journal of Mathematical Education in Science and Technology*, *53*(1), 11–34. https://doi.org/10.1080/0020739X.2021.1962554

Luneta, K. (2022). Special challenges in mathematics education in Sub Sahara Africa. *Current Opinion in Behavioral Sciences*, *48*, 101211. https://doi.org/10.1016/j.cobeha.2022.101211

Mukhanji, J. M., Ndiku, J. M., & Obaki, S. (2016). Effect of increased student enrolment on teaching and learning resources in Maseno University, Kenya. *The International Journal of Social Sciences and Humanities Invention*, *3*(3), 1938–1947. https://doi.org/10.18535/ijsshi/v3i3.05

Nicaud, J. F., Bouhineau, D., & Chaachoua, H. (2004). Mixing microworlds and CAS features in building computer systems that help students learn algebra. *International Journal of Computers for Mathematical Learning*, *9*(2), 169–211. https://doi.org/10.1023/B:IJCO.0000040890.20374.37

Oyengo, M. O., Parsons, D., Stern, D., & Sangwin, C. (2021). Providing student feedback through electronic assessment for linear algebra at Maseno University, Kenya. In *International meeting of the STACK community 2021*. Zenodo.

Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas*. Harper Collins.

Quigley, M. T. (1988). *Computer tutoring in mathematics education using artificial intelligence tools* [PhD, University of Birmingham].

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford University Press.

Sangwin, C. J., & Bickerton, R. (2023). Practical online assessment of mathematical proof. *International Journal of Mathematical Education in Science and Technology*, *53*(10), 2637–2660. https://doi.org/10.1080/0020739X.2021.1896813

Sangwin, C. J., & Grove, M. J. (2006). *STACK: Addressing the needs of the "neglected learners"*. Proceedings of the First WebALT Conference and Exhibition, January 5–6, Technical University of Eindhoven, Netherlands, 81–95.

Sangwin, C. J., & Kinnear, G. (2022). Coherently organised digital exercises and expositions. *PRIMUS*, *32*(8), 927–938. https://doi.org/10.1080/10511970.2021.199935214

Sleeman, D., & Brown, J. S. (Eds.). (1982). *Intelligent tutoring systems*. Academic Press.

UNESCO. (2010, December). *Trends in tertiary education: Sub-Saharan Africa* (No. 10). Unesco Institute for Statistics. UIS Fact Sheet.

Watters, A. (2021). *Teaching machines: The history of personalized learning*. MIT Press.

13

STUDY APPROACHES IN INTERACTIVE LEARNING ENVIRONMENTS FOR FORMATIVE ASSESSMENT

Detecting patterns aligned with Entwistle's theory

Maria Margeti and Manolis Mavrikis

1 Introduction

As intelligent interactive learning environments (ILEs) are becoming more prevalent in education, it is necessary to understand better how learners interact with them in order both to support their learning directly and to provide more information to their educators. The need for such efforts has been particularly evident, for example, in recent UNESCO guidelines that emphasise the importance of co-designing intelligent ILEs through collaboration between teachers, learners, and researchers (UNESCO, 2023). An example of the application of these recommendations is demonstrated in the development and evaluation of Google's generative AI tutor LearnLM-Tutor, in which the authors conduct co-design activities and apply metrics for the evaluation of AI tutors based on existing pedagogical frameworks (Jurenka et al., 2024).

In this chapter, we rely on the theory behind Entwistle's Approaches and Study Skills Inventory for Students (ASSIST; Entwistle, 1997), as it can support us in designing sophisticated ILEs. Such design would emphasise a deeper, more exploratory, and meaningful approach to learning that is particularly relevant when ILEs are used for practice exercises in the classroom in a form of formative assessment designed to support Black and Wiliam's key process in learning and teaching, such as establishing where the learners are in their learning, where they are going, and what needs to be done to get them there (Black & Wiliam, 2010). The intention is therefore to develop metrics that can identify whether students are adopting a deep approach to learning (seeking thorough understanding) or a surface approach to learning (focusing on minimal engagement). This distinction is particularly important because, as we

DOI: 10.4324/9781003533764-13

explain in more detail in Section 2.4, previous research suggests that adoption of deep approaches to studying correlates positively to performance, in contrast to a surface approach that correlates negatively to performance (Entwistle & Ramsden, 1983; Herrmann et al., 2017; Tait & Entwistle, 1996). In particular, surface approach is considered a good predictor of academic achievement (Diseth & Martinsen, 2003).

The study aims to give us insights on how students' approach to learning is linked to their use of an ILE for a specific learning scenario. Due to word limit, in the current chapter we only discuss in detail the hypotheses and model regarding the relationship between the surface approach to learning, as measured through ASSIST, and the specific metrics that emerge from students' usage of an ILE for mathematics when practicing exercises during tutorial sessions in the classroom.

2 Methodology

2.1 ActiveMath and the design process

ActiveMath (AM) is an interactive learning environment (ILE) which presents mathematical concepts, procedures, examples, and exercises in verbal (text and numerical) and visual form and allows to keep logs of users' actions via numerous metrics of usage. Its design was customised to serve the research aims of the current study, but also the requirements of the mathematics course, as the study was conducted in real learning conditions.

First, the AM interface and features were customised to integrate the learning content as defined by the official syllabus and the learning outcomes of the specific university course. Throughout the design process, storyboards were designed incorporating the feedback given by the mathematics team and seeking the team's approval before proceeding with the implementation of AM. The researcher (the first author) involved in this process needed to keep a fine balance between following what was officially approved by the university's processes and regulations as the mathematics curriculum for the course and providing more engaging and interactive ways for the students to practise their exercises in class. For example, while it was agreed with the mathematics team that it was important to keep consistency between the official textbook used for the module and the way the exercises are presented in AM as much as possible (so its integration would not trigger any course validation process), the mathematics team and the researcher involved also identified the need to incorporate features, such as a graph plotter, to improve visualisation and representational fluency and potentially boost students' comprehension, motivation, and engagement (Borba et al., 2013; Engelbrecht & Harding, 2005a, 2005b; Joshi, 2017; Juan et al., 2008; McDonald & Stevenson, 1999; Taleb et al., 2015).

Second, AM's design was customised so it could serve methodological aspects. For example, the learning content was structured in AM so that it facilitates the data collection in terms of the metrics of usage. There was a clear distinction between 'reading pages' and 'exercise pages' so it was possible, for example, to distinguish between temporal metrics (and consequently data) for 'reading pages' and 'exercise pages' per student.

This design process resulted in the following main features:

- 'Reading pages' allow students to view concepts, theory, and working examples for mathematical procedures. In the text of these pages, students could also use hyperlinks to other concepts (which can be either a revisitation of a previous concept or an introduction to a new one).
- 'Exercise pages' contain groups of exercises, allowing students to select which exercise to work on, such as multiple choice, multiple selections, and fill-in-the-blank. After submitting their answer, students receive feedback on whether their answer is correct. If the answer is incorrect, then students are encouraged to try again up to three times (which is a requirement set by the mathematics team). Throughout their three attempts, the previous wrong answers are retained and highlighted by AM. After the third attempt, they receive the correct answer. Students can also cancel an exercise at any point.
- Students can annotate private or public notes for each exercise through the feature of 'interactive notes', an editable notepad which encourages students to articulate their findings (Crowe & Zand, 2000; Engelbrecht & Harding, 2005a, 2005b; Galbraith & Haines, 1998; Laurillard, 2002; Sangwin, 2004).

Finally, it is worth noting that this study examines how students' naturally approach learning. Because of this, some other features of AM that might have provided guidance or support to the students were not activated in this study. The system functioned as a self-directed learning environment where students controlled their progress through theory, examples, and exercises, seeking tutor assistance when needed. In this way, the students' interactions with the ILE were captured without any additional interventions, which was part of the study's plan, which is described in more detail in the following.

2.2 Description of the study

The sample consists of 233 undergraduate students from a UK-based university who attended at the common core module of mathematics in the course of computer science that took place during the first semester of the first year of their studies in said UK university. In the sample, there are 190 males and 43 females, aged 18 to 46 years.

The study followed a non-experimental type of design, which allowed the study of learning behaviour in a more realistic setting, without manipulation of variables, and the examination of a plethora of variables and their relationships (Cohen et al., 2000; Robson & McCartan, 2016). It was carried out over the course of four weeks under real learning conditions, during the two-hour weekly tutorials of the module. During the first two weeks, students were introduced to and registered in AM. Each student had their own profile account set up. During the third and fourth weeks, AM was used for the first time for learning purposes, specifically for the chapters of 'functions-graphs' and 'matrices', respectively. It is also worth mentioning that AM was used in a blended learning mode – meaning, that it was used both during the tutorials for practising exercises and as well as after the face-to-face, two-hour lecture for that specific week. Students could also use AM's 'reading pages' during the tutorials to revisit concepts and theory explained in the lecture. They also had further assistance from the tutor assigned to each tutorial. Prior to the start of the semester, all tutors were shown AM, had accounts created for them, and had opportunities to discuss with the researcher how to integrate the ILE in the tutorials.

Data collection was conducted as follows: (1) AM tracked users' actions, collecting raw data based on which the interaction metrics were developed (see Section 2.5); (2) ASSIST was administered to students, who completed it voluntarily; (3) data in relation to prior knowledge (that is, their mathematics level based on their entry qualifications), gender, and age were also collected with the completion of the ASSIST measurement; and (4) observations were conducted during the tutorials to observe, for example, the extent to which students needed tutor intervention while working on their exercises. Finally, all the required ethical issues were addressed and applied throughout the planning and conducting of the study (Margeti, 2018).

2.3 Research questions

Based on the original study by Margeti (2018), the research questions for the study presented in this chapter were as follows:

1. What is the relationship between students' interactions in the learning environment AM and the deep and surface approaches towards studying when practising mathematics exercises during tutorial sessions in the context of formative assessment?
2. To what extent do students' interactions in the learning environment AM explain deep and surface approaches when practising mathematics exercises in tutorial sessions?

2.4 Selection of Entwistle's ASSIST

After conducting a thorough literature review of theoretical constructs related to different learning approaches (see Margeti, 2018), Entwistle's ASSIST (1997) was found to be the appropriate choice as a well-established, empirically validated framework for distinguishing studying approaches. ASSIST is a questionnaire that measures students' learning approaches through a series of items, categorising their tendencies into three main scales: deep, surface, and strategic. According to Entwistle (1997), 'deep' approach indicates students who seek to understand thoroughly and make meaningful connections. The 'surface' approach regards the extent to which students have the intention to complete a task with very little engagement, prefer unrelated memorising, cope minimally with the course, follow strictly the instructions and structure of learning content, and are motivated by 'fear of failure' (Entwistle, 1997; Entwistle & Ramsden, 1983). The 'strategic' approach describes students who adapt their learning methods to strategically achieve the highest possible grades based on what approach they perceive will be most effective.

In more detail, the reasons for the choice of ASSIST were:

1. The rationale behind its development suits this study's real learning conditions. Based on interviews and observations in real-life contexts in higher education, it gives an authentic account of learning context in higher education and provides the potential to capture and identify any complexities involved in how students interact in a learning environment in natural settings.
2. Its varied educational research background shows that it can capture the complexity of studying in different environments without oversimplifying by simply labelling students independently of the learning environment, task, or subject area. It acknowledges that the sustainability of a studying approach depends on the academic task, subject area, and learning environment.
3. Its construct validity and internal consistency were considered good based on independent evaluations, but there is a need for independent evaluation of its reliability (Coffield et al., 2004). Furthermore, in terms of internal reliability, using Cronbach's α, the deep and surface main scales were tested and had good internal consistency, with $\alpha = 0.804$ and $\alpha = 0.837$, respectively.

2.5 Metrics of usage in AM learning environment

Previous studies looking at the relationship between approaches to learning and students' interaction when using ILEs have examined metrics such as the number of pages visited and revisited, the time they spend on the system, and the use of navigational options that allow them to have a more global or detailed view of the learning material (Chen & Macredie, 2002; Mimirinis & Dafoulas, 2008). In the context of the current study, a 'revisitation' metric, such as the *relative amount of revisits*, (the probability that the visit to any URL is a repeat of a previous visit) (Tauscher & Greenberg, 1997), may predict students with a surface 'unrelated memorising' approach, who tend to rehearse and overlearn the learning content (Entwistle, 1997, 2001; Entwistle & Ramsden, 1983).

In addition, based on the research work of Botafogo et al. (1992), Herder and Juvina (2004), Juvina and van Oostendorp (2006), and McEneaney (2001), we included the 'path' metrics of *stratum* and *compactness*, which reveal the extent to which students follow the given structure of AM's table of content and a compact path (i.e. interact more closely around a certain set of AM pages), respectively. In the context of the current study, students with an intention for 'unrelated memorising' with tendencies for repetitive overlearning (Entwistle & Ramsden, 1983) are likely to interact more closely around a certain set of pages, which can result in high values in *compactness*.

However, the current study did not rely only on existing metrics from the existing literature at the time the study was planned but also expanded on metrics which can potentially be linked, initially theoretically, and then empirically, to ASSIST's studying approaches. As the main activity in AM was practising exercises, 'performance' metrics were developed capturing data on students' responses to these exercises. These metrics can indicate whether students get the correct answer; whether they get it correct on the first, second, or third attempt; whether they do not manage to solve a question at all; or whether they cancel before solving an exercise. As Entwistle and Ramsden (1983) find that scores of ASSIST's surface scale have negative correlation with performance, data occurring from such metrics can potentially reveal relationships with other surface subscales.

Hence, there was an initial selection as to the metrics that would be included in the multiple regression models.[1] They were selected according to whether they enable us to capture how students move around and what students do in AM, and whether the metrics can have empirical and conceptual connections to the ASSIST deep and surface scales and subscales. In Section 2.4, there is further discussion into the hypothesised relationships between the ASSIST surface scale and the AM metrics, as in this chapter we will focus on the surface approach. Finally, in Section 3 it is indicated which metrics from the initial selection process 'survived' in the final multiple regression model developed for the ASSIST surface scale (i.e. which metrics best predict a surface approach towards studying).

2.6 The ASSIST instrument and hypothesised relationships with AM metrics

We employed the second part of the ASSIST questionnaire that provides a student's 'approaches to studying' and consists of 52 items (Entwistle, 1997). In this chapter we focus on the surface approach and investigate its relationship to AM 'interaction' metrics with the intention of forming hypotheses and for which we provide some examples.[2]

As students may perceive a learning environment as threatening (Fransson, 1978), we developed the hypothesis that fear of failing may have a negative impact on students' performance when practising their exercises. In AM students can evaluate their answers and receive feedback. If the answer is incorrect, they are encouraged to try again and are allowed three attempts in all. As indicated in similar systems (Baker et al., 2008; Mavrikis, 2010), students' interaction with the exercises can indicate whether students are simply taking advantage of the affordances of the environment to achieve good results but with questionable learning gains. Students with a high degree of fear of failure may well exhibit 'gaming behaviour', because it allows them to succeed without risking failure (Entwistle et al., 1979). In addition, studies with first year undergraduates show that there is a statistically significant, negative correlation between the surface approach and performance, as mentioned in Section 1. This leads us to the general assumption that students with high surface scores are not likely to perform well in our context. We expect a negative association between surface scale and the *number of exercises solved on the first attempt*, and positive associations between fear of failure subscale and the metrics of *number of exercises solved on the second/third attempt* or *number of exercises finished but not solved*.

Furthermore, students with a high score on the surface scale may work slower, putting more effort into tasks and persisting longer when solving exercises due to their anxiety (Entwistle, 1981). So it is expected that students with high scores on the surface scale may result in high values on *average view time on exercise and reading pages*. The metrics of *maximum view time on exercise and reading pages* can also indicate some sort of 'extreme temporal interaction' or imbalance in terms of how students allocate their time when visiting reading and exercise pages due to the aforementioned difficulties, resulting in a positive association with the surface scale.

Finally, based on what is discussed in Section 2.3, we also expect a positive association between surface scale and the metric of compactness.

2.7 Strategy of analysis

The analysis concerns the usage of AM in which it was used for learning purposes as planned (during the third and fourth week of the study). The analysis was carried out on 115 students of the sample (out of the 223 described in Section 2.1) who attended and used AM during the third and fourth weeks of the study, and who also completed ASSIST (and after the process of data cleaning).

To examine the relationships between deep and surface approaches and 'interaction' metrics but also go a step further and examine which combinations of metrics can explain or predict deep and surface approaches, correlational analysis and multiple regression analysis are conducted.

Applying correlational analysis prior to multiple regression is quite common, but it is especially useful for a 'cold start' type of investigation, such as the current one, where there are really no prior empirical findings in a similar context to indicate which metrics are the most relevant for each approach. Furthermore, ASSIST is an ordinal-level measurement, whereas the students' 'interaction' metrics are considered interval-level measure; hence, a Spearman correlation coefficient (r_s) is the most appropriate statistic to examine this relationship.

The strategy with regards to the development of regression models is formed based on a combination of the accepted statistical practices when developing multiple regression models (Field, 2009) and specific tactics which have been adopted to aid and enrich the interpretation of the empirical findings in the specific context. A 'research-controlled' method was conducted where the researcher judges which predictors should be included in the model (Field, 2009; Meyers & Gamst, 2013). Prior to selecting the initial predictors (i.e. metrics) for the model of the ASSIST scale, we also considered the commonly used thresholds to judge multicollinearity between predictors (Meyers et al., 2006).

Finally, Cohen (1992) indicates that to ensure statistical power of 0.8 (a commonly accepted threshold) and if, with regards to the regression models, the statistical significance is set at $a = 0.05$[3] and the expected effect size is a medium $f^2 = 0.15$,[4] then there should be sample for at least 107 participants corresponding to eight predictors. Hence, it was decided that a maximum of eight predictors (i.e. metrics) is reasonable for the current study's sample of 115 participants.

3 Results

In the following analysis, we will focus on the regression model that predicts the surface scale, indicating at the same time the strategy we followed in the development of the regression models for the rest of the study approaches.

3.1 Initial selection of predictors

We start with the initial selection of metrics which are likely to be the best predictors for the regression model of the surface scale, after conducting the correlational analysis. We took advantage of the maximum number of predictors allowed (see Section 2.6), and we included eight metrics, as shown in Table 13.1.

TABLE 13.1 Predictors and reasons for including them in the initial regression model

Selected predictors/metrics	*Reason for selection*
Number of exercises solved on first attempt (r_s = –0.368)* *Number of exercises solved on third attempt (r_s = 0.314*)* *Number of exercises finished but not solved (r_s = 0.270*)* *Compactness (r_s = 0.216*)* *Average view time on exercise pages (r_s = 0.183*)*	• Statistical (based on statistically significant correlations with the surface scale) • Theoretical connections and potentially enriching the discussion further (see 2.4 and 2.5)
Number of hyperlinks (concept links) visited in reading and exercise pages (r_s = 0.197)*	• Statistical (based on statistically significant correlations with the surface scale)
Maximum view time on exercise page *Maximum view time on reading page*	• Theoretical connections and potentially enriching the discussion further (see 2.4 and 2.5)

**P <0.05.*

The direction of the statistical correlations, although they range from low to moderate in terms of practical significance, was as expected in the assumptions made in Sections 2.4 and 2.5.[5]

3.1.1 Development of regression model

After conducting multiple regression with the predictors shown in Table 13.1, we got the initial model, in which the variance explained R^2 was 37.3% and the adjusted R^2 was 32.5%, and with overall Sig. = 0.000. The regression was re-run several times to increase the variance and bring closer the values of R^2 and adjusted R^2. To achieve this, exclusion of both outliers and predictors[6] was conducted. This resulted in the final model with seven predictors (those shown in Table 13.1, except *average view time on exercise pages*). This model explained a large amount of variance with R^2 at 45.5% and the adjusted R^2 at 41.8%. Given that a medium amount of variance was expected, this is an indication that the model is generally giving us a good picture of students' interaction in the AM with regards to the surface approach towards studying. The final model is overall statistically significant (Sig. = 000) and holds well with regards to the assumptions typically checked in regression models (see Figure 13.1). These results give us reasonable confidence in terms of the generalisability of results in the same or similar contexts, although there are limitations, as discussed later.

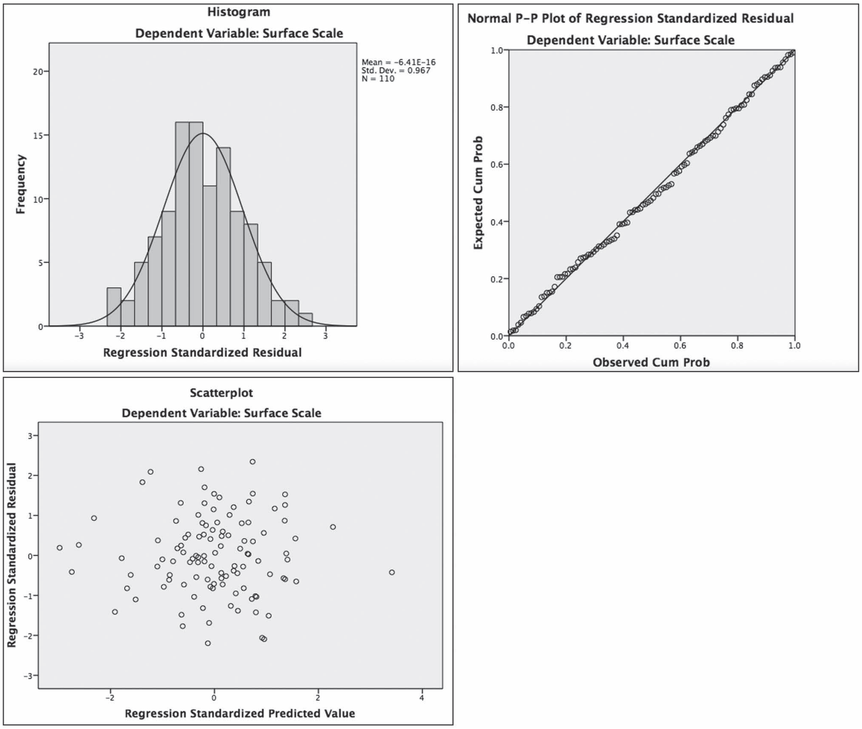


FIGURE 13.1 (a) Histogram for the assumption about the normal distribution of residuals, (b) plot for the normality assumption, and (c) scatterplot for the assumption about the homoscedasticity of residuals.

3.2 The influence of prior knowledge in the surface regression model

The final surface model explained a large amount of variance, but there is variance which is unexplained. One reason can be the influence of prior knowledge. Osmon (2009) notes that the struggle with mathematics during the first year of study in computer science courses is a widespread problem and notes that students' mathematics qualifications on entry to university computer science courses may be one of the reasons. Furthermore, prior knowledge can influence the way students interact with an ILE and their studying approach (Chen & Ford, 2000; Chen et al., 2016; Chen & Paul, 2003; Ramsden, 2005).

Hence, starting from the final version of the surface model, prior knowledge is included as a 'selection' variable when running the multiple regression in SPSS and the sample is split into a 'low prior knowledge' group and a 'high prior knowledge' group.[7] After running the multiple regression for the surface model with the split groups, the variance explained by the model for the 'low prior knowledge' group increased with R^2 = 55.8% and adjusted R^2 = 47.2%, while for the 'high prior knowledge' group it decreased slightly, with R^2 = 40.7% and adjusted R^2 = 33.6%. The surface models for both groups were overall statistically significant.

4 Discussion

4.1 Reflections on the surface regression model

The results suggest that aspects of students' usage in the mathematical learning environment AM can relate to the studying approach. The final model with its seven predictors can enrich the interpretation of findings, as it can draw a pretty good picture of (and help identify) students with high scores on the scale. Specifically, if tutors (human or AI ones) detect that students repeatedly do not manage to solve the exercises on first attempt and need to try several times (indicating 'gaming the system'), then there is cause for intervention. This is especially necessary if exercises with the same level of difficulty or type continue to be solved on subsequent attempts or not at all. Furthermore, the *maximum view time on an exercise page* increases as the surface score increases, which can be interpreted as 'getting stuck' on a specific exercise, triggering again a tutor intervention. Finally, students with high scores on the scale tend to follow a rather compact limited path when going through the learning material AM. On the other hand, there are unexpected relationships. With regards to *maximum amount of time on a content (reading) page*, the direction of its relationship to surface scale is negative (*beta* = *–0.188*). It seems that, in the context of the current study, students with a high score on the surface scale are less likely to spend an increasing amount of time on a specific theoretical page, compared to those with low scores. This is also reinforced by the observations made in class, as some students would not go through the theory, even when they were 'getting stuck' on specific exercises, unless the tutor advised them to do so. This can be a sign of lack of self-regulation, which in other studies has been related to surface scale (Lindblom-Ylänne et al., 2019).

The influence of prior knowledge in the model indicates that it is a factor that should be included in regression models that express approaches towards studying, especially when students are practicing tutorial exercises.

The combination of well-selected predictors and level of prior knowledge can help flag up a surface approach if incorporated in an intelligent system and help both human and AI tutors intervene and provide appropriate feedback during formative assessment.

4.2 Expanding the process to all ASSIST scales and ideas on serving ILEs and AI tools

The research work presented here is part of a bigger research project in which the same strategy was applied for the deep and surface scales and their eight subscales of ASSIST, with and without the influence of prior knowledge. From these results, relevant to the current discussion are the models of the surface subscales *fear of failure* and *unrelated memorising*, with variance at 41.6% and 40%, respectively; the model of deep scale, with variance at 35.1%; and the models of deep subscales *relating ideas* and *seeking meaning* in the 'low prior knowledge' group, with variance at 43.1% and 41.3%, respectively (Margeti, 2018). Each model was explained by a different combination of predictors/metrics. All models rated from moderate to large variance and had overall statistical significance and hold well with the typical assumptions for regression models (see Margeti, 2018). Given the study's real conditions, the overall results give us reasonable confidence to examine ways the detection of studying approaches through interaction metrics may serve ILEs.

Firstly, detecting studying approaches may have implications for the design of an ILE by clarifying whether features which students are using when working on tutorial mathematics exercises promote a specific approach. For example, there is the potential to reveal features in an ILE whose use promotes a surface approach rather than the deep approach. This is the case of AM's concept links (i.e. hyperlinks), which was found to be a positive predictor for the surface scale model (beta = 0.101), while it was a negative predictor for the deep subscales models of relating ideas (beta = −0.197) and seeking meaning (beta = −0.138) (Margeti, 2018). Hence, the use of hyperlinks is not necessarily linked to the deep approaches, as initially assumed, and they are even used by those with higher scores on the 'surface' scale. This indicates that the design and the content of hyperlinks require changes, so instead of contributing to repetitive overlearning or unreflective memorisation, they will aim to extend further what is currently known (Margeti, 2018). The findings trigger also the need for the introduction and integration of more sophisticated ILE features which facilitate links between mathematical concepts and understanding of their relationships when students are working on exercises and encourage exploration, such as concept maps. These uses may encourage students to seek the underlying meaning of what they learn (i.e. seeking meaning) and explore the relationships between concepts with the intention of deeper understanding (i.e. relate ideas) (Entwistle & Nisbet, 2013; Entwistle & Peterson, 2004) when working on their tutorial exercises.

Secondly, by being able to detect a studying approach through interaction metrics, it may be possible to translate these connections into measurable tutor behaviours by working, for example, with practitioners and experts in studying approaches, creating in this way pedagogical rubrics for the development and evaluation of language models for AI tutors. This is in line with the methodology used by Jurenka et al. (2024), according to which they used pedagogical rubrics as part of a pedagogical development and evaluation framework for AI tutors. Jurenka et al. (2024) point out that this was triggered by the difficulties of 'verbalizing pedagogical intuitions into generative AI prompts' and the need for better evaluation practices.

Entwistle's ASSIST is a construct whose development is spanning over four decades in real contexts in higher education, and it continues to develop with new, robust updated inventories which are tested in a variety of learning scenarios (Lindblom-Ylänne et al., 2019). Hence, ASSIST studying approaches as pedagogical dimensions with their corresponding interaction metrics can provide a foundation for creating measurable tutor behaviours in essence metrics for language model evaluations (LME). It opens an opportunity for developing a methodology to create pedagogical rubrics according to which a language model (LM) is prompted to generate instructions on a given math task and topic based on a nuanced pedagogical framework and construct such as ASSIST. Table 13.2 gives an idea of how such a pedagogical rubric based on ASSIST can be formed.

The findings of the current investigation are robust enough to give confidence regarding the surface scales and their predictors, shown in Table 13.2, in this type of learning scenario and educational setting; however, for the current methodology to transcend in the context of LM development and evaluation, this would be only the start. Further investigation is required, particularly with regards to the 'tutor guidance' indicated in Table 13.2. The tutor guidance is currently based on the existing literature in studying approaches and the interpretation of this study's models (see Margeti, 2018). In the case of an AI tool, a complementary, more nuanced, qualitative approach would also be required, using interviews with both students and practitioners to define better the pedagogical rubric and ultimately enrich the AI prompts with tutor guidance, which promotes or discourages a specific approach in such learning scenarios. The participatory design process with practitioners and students followed by Jurenka et al. (2024) is an indication that such qualitative research methods are required for the formation of pedagogical rubrics.

5 Conclusions

During the current investigation, it has become apparent that Entwistle's instrument reflects an educational philosophy that recognises subtleties and offers a realistic view of a student's intentions and actions in a learning environment. At the time the study was planned, there was no similar research to uncover relationships between the students' approaches towards studying and students' interaction with ILEs during tutorial sessions in class. The findings presented in this chapter indicate that it is possible to identify tendencies in

TABLE 13.2 Pedagogical rubric based on surface scales with the interaction metrics/predictors of their models

Pedagogical dimension/ ASSIST scale	*Interaction metric/ predictor of models*	*Tutor guidance and potential basis for LME metric*
Surface	• *Number of exercises solved on first attempt (–)* • *Number of exercises solved on third attempt (+)* • *Number of exercises finished but not solved (+)* • *Maximum view time on an exercise page (+)* • *Maximum view time on a reading page (–)* • *Number of hyperlinks (concepts links) visited in reading and exercises pages (+)* • *Compactness (+)*	• Discourage abusing the ILE's affordances and exhibiting 'gaming' behaviour when practicing exercises. • Encourage connecting old concepts and relationships from students' prior knowledge to new ones to enhance understanding when students get stuck in specific exercises. • Encourage a less compact learning path by seeking further relevant working mathematics examples and gradually expanding to exercises of higher level of difficulty. • Encourage more advanced techniques for deeper understanding, like generating concept maps or exploring existing ones (especially if usage of system's hyperlinks shows that it promotes a surface approach like here).
Unrelated memorising	• *Number of exercises solved on first attempt (–)* • *Number of exercises finished but not solved (+)* • *Compactness (+)* • *Average number of times a 'notes' link is clicked per page) (+)*	• Discourage abusing the ILE's affordances and exhibiting 'gaming' behaviour when practicing exercises. • Encourage a less compact learning path by seeking further relevant working mathematics examples and gradually expanding to exercises of higher level of difficulty. • Encourage the use of the 'notes' feature for more advanced studying techniques to promote deeper understanding, like drawing concept maps (especially if the usage of the 'notes' feature shows that it promotes a surface approach like here).
Fear of failure	• *Number of exercises solved on first attempt (–)* • *Number of exercises solved on third attempt (+)* • *Maximum view time on an exercise page (+)* • *Maximum view time on a reading page (–)*	• Discourage abusing the ILE's affordances and exhibiting 'gaming' behaviour when practicing exercises. • Encourage connecting old concepts and relationships from students' prior knowledge to new ones to enhance understanding when students get stuck in specific exercises.

Note: Corresponding potential tutor guidance which can be the basis for LME metrics.*Source:* As developed in Margeti (2018).[8]

students' interactions in an ILE for formative assessment, such as AM, which are linked to the surface approach.

While this study provides valuable insights into the detection of studying approaches in this specific learning environment and sample, which was one of convenience, there are limitations in terms of generalisation (although the sample size was adequate, as indicated in Section 3.2). The study's real conditions provide ecological validity, but they do come at the expense of the study's internal validity.

Despite the limitations, this current investigation can provide a good starting point for future studies in different educational settings. First, researchers of such studies should consider the latest versions of ASSIST's instrument and make the necessary adjustments or introduce new 'interaction' metrics (performance, temporal, revisitation, or path). Second, that the study introduces the importance of prior knowledge as a factor in regression models highlights the necessity of modelling students when designing (or evaluating) formative assessment tools in mathematics. Third, the observed lack of self-regulation among high surface students underscores the importance of fostering self-regulatory skills (both from teachers and from AI-driven tutors) when students are working on mathematics exercises. Fourth, our approach can indicate whether the ILE in question includes features which promote a surface approach rather than a deep one and initiate changes in an ILE's design. Finally, this study opens the possibility of introducing Entwistle's approach in formative assessment tools in mathematics by flagging surface intentions and formulating relevant tutor guidance.

Acknowledgements

We would like to acknowledge the ActiveMath group for the help and support in pre-processing the log files and implementing the design changes, Harvey Mellar for supervising part of the work, and Vasos Pavlikas for providing the math content for AM and input with regards to its integration in AM. MM's contribution to this chapter was partly supported by the CHAILD project - Children's Agency In the age of AI: Leveraging Inter Disciplinarity (UKRI ref. MR/Z505882/1). The views expressed in this chapter are those of the authors and do not reflect those of UK Research and Innovation (UKRI).

Notes

1 See Margeti (2018) for a full discussion of the metrics used as predictors for the regression models of all the ASSIST deep and surface scales. In this chapter, we focus on the metrics used as predictors for the surface scale only.

2 See Margeti (2018) for a full discussion of all hypothesised relationships.
3 This is the expected level of significance, and all regression models will be developed according to this, which is the default in SPSS.
4 This is the expected effect size, based on the literature review conducted at the time of planning the study (see Margeti, 2018).
5 The exception is the correlation of the scale to *number of hyperlinks (concepts links) visited in reading and exercises pages*, which was more of a 'cold start' type of investigation.
6 The exclusion of predictors was based on their *beta* and *Sig*. values (Field, 2009).
7 Prior knowledge in mathematics was measured through students' entry qualifications, as indicated in Section 2.1, and the mathematics team advised on how to split the sample into the two groups (see further discussion in Margeti (2018)).
8 *The signs (+)* and *(–) next to predictors indicate the direction of their relationship to the surface scale and subscales based on the beta values of models.*

Reference list

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Black, P. J., & Wiliam, D. (2010). Inside the black box raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Borba, M., Clarkson, P., & Gadanidis, G. (2013). Learning with the use of the Internet. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international hanbook of mathematics education* (pp. 691–720). Springer.

Botafogo, R., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, *10*(2), 142–180.

Chen, S., & Ford, N. (2000). Individual differences, hypermedia navigation and Learning: An empirical study. *Journal of Educational Multimedia and Hypermedia*, *9*(4), 281–311.

Chen, S., Huang, P.-R., Shih, Y.-C., & Chang, L.-P. (2016). Investigation of multiple human factors in personalised learning. *Interactive Learning Environments*, *24*(1), 119–141.

Chen, S., & Macredie, R. (2002). Cognitive styles and hypermedia navigation: Development of a learning model. *Journal of the American Society for Information Science and Technology*, *53*(1), 3–15.

Chen, S., & Paul, R. (2003). Editorial: Individual differences in web-based instruction – an overview. *British Journal of Education Technology*, *34*(4), 385–392.

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning. A systematic and critical review*. Learning and Skills Research Centre.

Cohen, J. (1992). Quantitative methods in psychology – a power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. RoutledgeFalmer.

Crowe, D., & Zand, H. (2000). Computers and undergraduate mathematics 3: Internet resources. *Computer & Education*, *35*, 123–147.

Diseth, A., & Martinsen, O. (2003). Approaches to learning, cognitive style, and motives as predictors of academic achievement. *Educational Psychology*, *2*.

Engelbrecht, J., & Harding, A. (2005a). Teaching undergraduate mathematics on the Internet – Part 1: Technologies and taxonomy. *Educational Studies in Mathematics*, *58*(2), 235–252.

Engelbrecht, J., & Harding, A. (2005b). Teaching undergraduate mathematics on the Internet – Part 2: Attributes and possibilities. *Educational Studies in Mathematics*, *58*(2), 253–276.

Entwistle, N. (1981). *Styles of learning and teaching- an integrated outline of educational psychology*. John Wiley and Sons.

Entwistle, N. (1997). *Approaches and study skills inventory for students (ASSIST)*. ETL Project. Retrieved January 7, 2013, from www.etl.tla.ed.ac.uk/questionnaires/ASSIST.pdf

Entwistle, N. (2001). Styles of learning and approaches to studying in higher education. *Kybernetes*, *30*(5–6), 593–603.

Entwistle, N., Hanley, M., & Hounsell, D. (1979). Identifying distinctive approaches to studying. *Higher Education*, *8*, 365–389.

Entwistle, N., & Nisbet, J. (2013). The nature and experience of academic understanding. *Psychology of Education Review*, *37*(1).

Entwistle, N., & Peterson, E. (2004). Conceptions of learning and knolwege in higher education: Relationships with study behaviour and influences of learning environments. *International Journal of Educational Research*, *41*, 407–428.

Entwistle, N., & Ramsden, P. (1983). *Understanding students learning*. Nichols.

Field, A. (2009). *Discovering statistics using SPSS*. SAGE Publications Ltd.

Fransson, A. (1978). *Test anxiety and motivation to learn*. Acta Univertsitatis Gothoburgensis.

Galbraith, P., & Haines, C. (1998). Disentangling the nexus: Attitudes to mathematics and technology in a computer learning environment. *Educational Studies in Mathematics*, *36*, 275–290.

Herder, E., & Juvina, I. (2004). *Discovery of individual user navigation styles*. Adaptive Hypermedia AH2004 Workshop on Individual Differences, Eindhoven.

Herrmann, K. J., McCuneb, V., & Bager-Elsborga, A. (2017). Approaches to learning as predictors of academic achievement: Results from a large scale, multi-level analysis. *Higher Education*, *7*(1).

Joshi, D. R. (2017). Influence of ICT in mathematics teaching. *International Journal for Innovative Research in Multidisciplinary Field*, *3*(1).

Juan, A., Huertas, A., Steegmann, C., Corcoles, C., & Serrat, C. (2008). Mathematical e-learning: State of the art and experiences at the Open University of Catalonia. *International Journal of Mathematical Education in Science and Technology*, *39*(4), 455–471.

Jurenka, I., Kunesch, M., McKee, K., Gillick, D., Zhu, S., Wiltberger, S., Milind Phal, S., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pîslar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., & Ko, W.-J. (2024). *Towards responsible development of generative AI for education: An evaluation-driven approach*. https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf

Juvina, I., & van Oostendorp, H. (2006). Individual differences and behavioral metrics involved in modeling web navigation. *Universal Access in the Information Society*, *4*(3), 258–269.

Laurillard, D. (2002). *Rethinking university teaching – a framework for the effective use of learning technologies*. RoutledgeFalmer.

Lindblom-Ylänne, S., Parpala, A., & Postareff, L. (2019). What constitutes the surface approach to learning in the light of new empirical evidence? *Studies in Higher Education*, *44*(12).

Margeti, M. (2018). *Explaining students' deep and surface approaches to studying through their interactions in a digital environment for mathematics*. University College London.
Mavrikis, M. (2010). Modelling student interactions in intelligent learning environments: Constructing Bayesian networks from data. *International Journal on Artificial Intelligence Tools*, *19*(6), 733–753.
McDonald, S., & Stevenson, R. (1999). Spatial versus conceptual maps as learning tools for hypertext. *Journal of Educational Multimedia and Hypermedia*, *8*(1), 43–64.
McEneaney, J. (2001). Graphic and numerical methods to assess navigation in hypertext. *International Journal of Human-Computer Studies*, *55*, 761–786.
Meyers, L., & Gamst, G. (2013). *Applied multivariate research design and interprentation*. Sage.
Meyers, L., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research – design and interpretation*. Sage.
Mimirinis, M., & Dafoulas, G. (2008). *Patterns of use of virtual learning environments and students' approaches to learning: A case study of undergraduate students*. Proceedings of World Conference on Educational Multimedia Hypermedia and Telecommunications, Vienna.
Osmon, P. (2009). *Post-16 maths and university courses: Numbers and subject interpretation*. Proceedings of the British Society for Research into Learning Mathematics, Loughborough.
Ramsden, P. (2005). The context of learning in academic departments. In *The experience of learning: Implications for teaching and studying in higher education*. University of Edinburgh, Centre for Teaching, Learning and Assessment.
Robson, C., & McCartan, K. (2016). *Real world research*. Wiley.
Sangwin, C. (2004). Section A – assessing mathematics automatically using computer algebra and the internet. *Teaching Mathematics and Its Applications*, *23*(1).
Tait, H., & Entwistle, N. (1996). Identifying students at risk through ineffective study strategies. *Higher Education*, *31*, 97–116.
Taleb, Z., Ahmadi, A., & Musavi, M. (2015). The effect of m-learning on mathematics learning. *Social and Behavioral Science*, *171*, 83–89.
Tauscher, L., & Greenberg, S. (1997). How people revisit web pages: Emperical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, *47*(1), 97–137.
UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.

14

INTRODUCING ATOMIC, RE-USABLE FEEDBACK FOR THE SEMI-AUTOMATED ASSESSMENT OF MATHEMATICS TASKS

Filip Moons

1 Introduction

Feedback is widely accepted as a powerful engine of learning processes (Hattie & Timperley, 2007). However, feedback in mathematics classrooms is often limited to evaluative feedback – meaning, that students get grades on tests, and sometimes mistakes are highlighted, but with no word of explanation (Knight, 2003). Many policy reports from different countries and regions have pointed to this flaw. Reasons reported included large class sizes and time constraints (Eurydice, 2021; Gibson et al., 2015). Indeed, the Eurydice report (2021) pointed out that 49% of all the teachers in the European Union indicated having too much assessment work, making it the second biggest complaint about the teaching profession, after having too many administrative tasks. The lack of feedback in mathematics education is also a returning observation in the yearly reports of the Flemish Education Inspectorate (2023):

> *Feedback in mathematics classrooms is often too focused on the product with insufficient attention to the reasoning or arithmetical errors underlying the mistakes.*
>
> *(Flemish Education Inspectorate, 2023, p. 65, own translation)*

Digital assessments offer a promising solution to the increasing demand for more feedback in mathematics education. Over the past few years, technology has significantly influenced various aspects of assessment, such as enabling the automated generation and grading of test items (e.g. Olsher et al., 2024; Pelkola et al., 2018). Despite these advancements, pen-and-paper

DOI: 10.4324/9781003533764-14

assignments still dominate mathematics classrooms for various reasons (Lemmo, 2023). In some cases, traditional methods are more practical, while in others, the digital environment may pose challenges, such as difficulty in inputting intermediate steps (Bokhove & Drijvers, 2010). Moreover, certain tasks are naturally better suited to being solved with paper-and-pencil (Threlfall et al., 2007), and digital assessments often emphasise lower-order skills, like procedural tasks (Hoogland & Tout, 2018). Additionally, fully automating the assessment of complex mathematical tasks, such as proofs, remains a significant challenge (Olsher et al., 2024). In other cases, the effort required to develop fully automated assessments may not be justified.

A nice feature of students' solutions to mathematics tasks is that they often exhibit systematic error patterns (Movshovitz-Hadar et al., 1987; Schnepper & McCoy, 2014) – meaning that different students often make similar mistakes. When writing feedback, this means teachers often have to repeat the same comments multiple times, leading to one of the main ideas of this chapter: pieces of feedback can often be re-used for multiple students. A system that facilitates this re-use helps eliminate such repetition, possibly allowing considerable time savings and improved feedback while enhancing student learning. We call this approach semi-automated, as teachers still assess, but the computer helps by saving and suggesting feedback previously re-used. The concept of atomic feedback has been invented to write re-usable feedback.

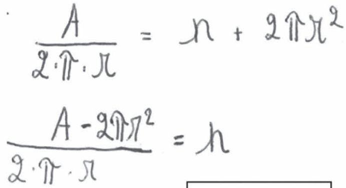
This chapter provides a comprehensive summary of the key findings from my PhD project, 'Semi-automated assessment of handwritten mathematics tasks' (Moons, 2023), while also offering a forward-looking perspective. The chapter begins by introducing the concept of atomic feedback, followed by a detailed presentation of a digital semi-automated tool designed to deliver this type of feedback. The subsequent section delves into the primary research results related to atomic feedback. Next, we explore two main applications: statement banks and checkbox grading. We conclude the chapter with a look at the future: How can we make suggestions for re-using feedback smarter?

2 Atomic feedback

Consider the classic feedback shown in Figure 14.1: While it is comprehensive and highly tailored to the student's specific answer, the likelihood of re-using parts of this text for other students with similar mistakes is relatively low. This is where the concept of atomic feedback comes into play. Atomic feedback provides a set of simple guidelines to follow when writing feedback on students' solutions to mathematics tasks to make them more re-usable (Moons et al., 2022), which will also be investigated in the following section.

Student's solution

Manipulate the formula: $A = 2\pi rh + 2\pi r^2$ to h



Classic feedback

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it shouldn't be placed directly in the numerator, as you should make the denominators the same.

Atomic feedback

- First step
  - Dominant operation on the right side is an addition!
    - Division of left-hand side is not helpful
    - $2\pi r$ is a common factor of the right side, but:
      - sum wasn't completely divided by it
      - the second addend was not divided
- Second-step
  - Your final answer is correct, but:
    - It looks like a coincidence.
    - You should subtract $2\pi r^2$ from both sides.
    - Mistake with making the denominators the same!
      - $2\pi r^2$ shouldn't be directly in the numerator.

FIGURE 14.1 Comparing atomic with classic feedback.

2.1 Definition of atomic feedback

To write *atomic feedback*, one must:

1. Identify independent errors
2. Write small feedback items for each error separately
3. If an error reflects a structural mistake or misconception, create two feedback items:
   a. One item addressing the misconception in general
   b. One or more sub-items targeting specific mistakes

Atomic feedback items eventually create a list of bullet points, focusing on the pertinent aspects of a student's solution. This list can be organised hierarchically to group related items together. Such clustering helps ensure that feedback is as atomic as possible and cuts overly specific feedback into separate items. Moreover, clustering offers a systematic approach to presenting related feedback to students, through thematic grouping or visually combining general and specific feedback on the same error.

Referring to Figure 14.1, we can observe that classic feedback can be rephrased into atomic feedback. It is important to remember that atomic feedback consists solely of form requirements without imposing content limitations. The listed items serve as examples of atomic feedback, with the hierarchy reflecting the steps in the student's solution and the types of errors made. The structural mistake related to making the denominators the same leads to two feedback items: one that highlights the structural error (related to making the denominators the same), and another that addresses the specific mistake ($2\pi r^2$ should not be directly in the numerator), complying to rule 3 in the definition.

2.2 Hypothetical division into sub-items

When determining whether a feedback item is atomic, one can attempt to divide it into sub-items. If this hypothetical division results in sub-items that make sense independently of the specific context of the task being corrected, the original item is non-atomic. For example, the feedback '*Neither the choice of the unknown nor the starting equation is correct*' can be split into two items:

- *Choice of the unknown: wrong*
- *Start of the equation: wrong*

2.3 Independent errors

Atomic feedback requires identifying independent errors and addressing each one in separate items. The previous example, '*Neither the choice of the unknown nor the starting equation is correct*', is a good example of an item that addresses two errors and, as such, violates independence. Independence can also be compromised if a feedback item refers to another item, making it unusable on its own. For example, a comment like '*Idem to comment above*' cannot stand alone without the 'comment above'.

2.4 Atomic or not?

The preceding guidelines offer a framework for writing atomic feedback across a wide range of mathematical subjects. However, determining the atomicity of feedback items for research purposes can be a complex task. Appendix D in Moons et al. (2022) offers a comprehensive codebook with concrete examples to distinguish between atomic and non-atomic feedback items.

3 A semi-automated digital tool for atomic feedback

A Moodle plug-in (Gamage et al., 2022) was developed to implement the concept of re-usable feedback. This tool allows teachers to create a hierarchical list of feedback items, which the system stores for future use when

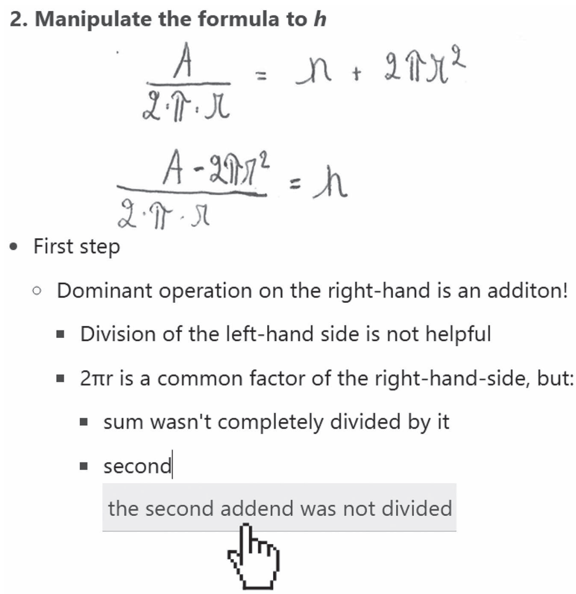


FIGURE 14.2 Screenshot of the digital SA tool.

relevant to another student. The tool suggests previously used feedback items by attempting to match them with what the teacher is currently typing. In the final section of this chapter, we will explore this non-intelligent suggestion system in greater detail, as it will turn out to be a crucial element in the re-usability of (atomic) feedback.

Figure 14.2 illustrates how the tool functions in practice. When providing feedback on a student's solution, teachers have three options: indicating that the solution is perfect, noting that an answer is missing, or providing atomic feedback. They can use keyboard shortcuts to create a hierarchical list of feedback items.

Although the digital tool was developed within the framework of atomic feedback, it is important to note that the end users ultimately determine whether their feedback adheres to the atomic feedback guidelines. Therefore, throughout the rest of this chapter, we will refer to this digital tool as the **SA** tool (semi-automated tool), reflecting the collaboration between teachers and the computer. This distinction allows us to separate the influence of the tool's characteristics from the influence of the feedback's atomicity.

4 The influence of atomic feedback on time investment, feedback quantity, re-usability, and quality

A cross-over experiment was conducted with 45 Flemish mathematics teachers to examine the effect of atomic feedback on time investment, feedback quantity, re-usability, and quality. The initial iteration of the experiment served as a pilot study involving 9 teachers, while the main study included data from 36 teachers.

The mathematics teachers were sampled using announcements in math teaching magazines: 28 female and 17 male mathematics teachers participated, with an average age of 40.2 years (SD = 10.3).

4.1 Study design

The experiment compared two conditions: the SA condition, where teachers used the SA tool and were encouraged to provide feedback in atomic units, and the **PP** condition, where teachers were given only a textbox to provide feedback, without the option to re-use previously entered feedback items, simulating a pen-and-paper approach (see Figure 14.3).

During a full working day, half of the teachers started in one condition and swapped to the other in the afternoon. Each teacher gave feedback on 60 real students' solutions to a linear equations task consisting of three questions: (1) solving an equation, (2) manipulating a formula (see Figures 14.1, 14.2, and 14.3), and (3) a word problem (see Figure 14.7). A quasi-random selection of 30 solutions was assessed in each condition. This quasi-randomness was essential to ensure: (1) *Comparability of the feedback between the SA and PP condition*. Each student's solution was included 18 times in the SA condition, and 18 times in the PP condition across all 36 teachers. (2) *Balanced conditions for each teacher:* 10 good, 10 moderate, and 10 bad students' solutions were included in each condition based on the pilot study data, ensuring that the SA and PP conditions were balanced. (3) *Random order.* The order in which these solutions were assessed in each condition was random.

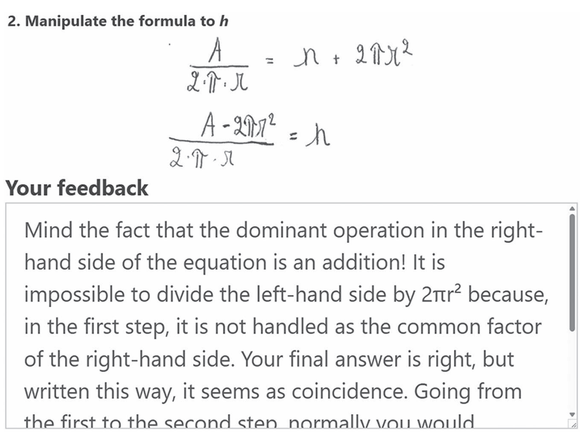


FIGURE 14.3 Screens of the PP condition.

4.2 The effect of atomic feedback on feedback re-usability

Atomic feedback was developed to provide guidelines for writing feedback that is more re-usable. However, does it work? One way to investigate this is by categorising all feedback items in the SA condition as either 'atomic' or 'non-atomic' and then examining whether there is a relationship between the item's atomicity and whether it is re-used or not.

In Moons et al. (2022), the 2,424 feedback items from the SA condition were coded according to this classification using the aforementioned codebook. It was found that 73.7% of the feedback items were atomic, while 26.3% were non-atomic. The most common violations of atomicity involved items that discussed multiple errors simultaneously (31.9% of all non-atomic items).

Next, the relationship between atomicity and the re-usability of feedback items was examined. An item was considered 're-used' if used more than once. The crosstabulation results are displayed in Table 14.1. A chi-square test of independence showed a significant relationship between these variables: Atomic items were significantly more likely than non-atomic items to be re-used, $\chi^2(1, n = 2{,}424) = 85.34$, $p < .001$.

4.3 Effect on time investment and feedback quantity

The goal of re-usability of feedback is to ultimately reduce the workload. Therefore, the total time spent on each task was measured in both conditions. Figure 14.4 shows the time spent assessing tasks in the SA and PP condition in function of the number of tasks assessed in each condition. On average, 3 minutes, 11 seconds were needed to give feedback to a task in the PP condition, and 3 minutes, 18 seconds in the SA condition. When performing a paired *t*-test on the total time spent in both conditions, no significant difference could be found between the SA and PP conditions; $t(35) = 0.002$, $p = .998$.

When comparing the feedback quantity, we compared the average number of characters that each teacher used in each condition per task (see the boxplots in Figure 14.5). A paired *t*-test revealed a significant increase in

TABLE 14.1 Crosstab between atomicity and re-usability of feedback items

	Re-usability		
Atomicity	*Re-used item*	*Non-reused item*	*Total*
Atomic item	731 (40.9%)	1,055 (59.1%)	**1,786 (73.7%)**
Non-atomic item	131 (20.5%)	507 (79.5%)	**638 (26.3%)**
Total	**862 (36.6%)**	**1,562 (64.4%)**	**2,424 (100%)**

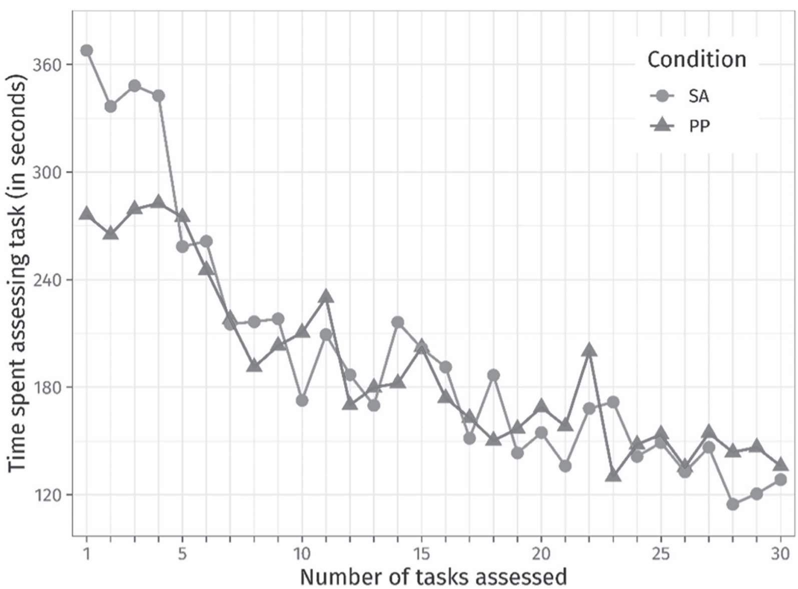


FIGURE 14.4 Evolution of teachers' time in both conditions as more and more tasks get assessed.

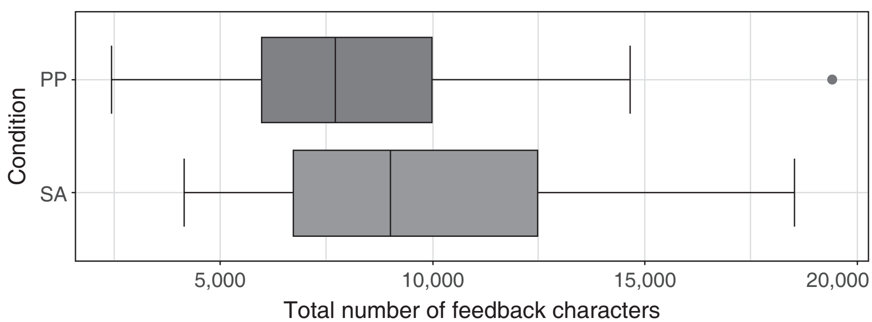


FIGURE 14.5 Boxplots of the total number of feedback characters teachers use in each condition.

the average number of characters in the SA condition (M = 9,656 chars, SD = 3,553 chars) relative to the PP condition (M = 8,409 chars, SD = 3,672 chars); $t(35) = 2.43$, $p = .02$. The effect size, expressed as Cohen's d, is 0.41, which approaches Cohen's threshold for a medium effect.

These findings suggest that rather than saving time, teachers provided significantly more feedback when using the SA tool. In other words, instead of completing the assessment process more quickly, the teachers offered more feedback. Some teachers even noticed this during the experiment, with one commenting, *'I think I'm giving much more feedback using your system'*. Although the sample predominantly consisted of highly motivated mathematics teachers, the result was unexpected, as the teachers were instructed to provide comparable feedback regarding the quantity, quality, and content in both conditions. This outcome also serves as a caution for those seeking to reduce the workload in the teaching profession. Although finding such solutions is critically important (Gibson et al., 2015), some teachers may perceive them as opportunities to increase their workload.

From a methodological standpoint, our results highlight the importance of being cautious about claims regarding time savings in assessment work when time developments are presented without any reference point. If we had only considered the SA graph in Figure 14.2, we might have concluded that SA leads to time savings as more tasks are getting assessed. However, without comparing these results to those in the PP condition, it is difficult to recognise that the observed downward trend is due to increasing task familiarity (Lim et al., 1996). While we found no studies explicitly examining task familiarity in the context of teachers' assessment work, it is highly plausible that teachers eventually memorise their solution keys and become more efficient in providing feedback.

4.4 Effect on feedback form, content, and quality

Moons et al. (2024) investigated the similarities and differences between the feedback on form, content, and quality between the SA and PP conditions. This was done using text mining and qualitatively coding all feedback reports in categories distilled from the literature on effective feedback in mathematics education (Busch et al., 2015b, 2015a). An important distinction from the previous results is that this study considers feedback *reports*: the whole feedback text given to a student's solution to a question. In the analysis of atomicity above, items from the SA condition were examined at the level of separate list entries rather than full reports. In this section, we focus on the main takeaways of this study.

The text mining analysis performed a word frequency analysis on the feedback reports in both conditions (Silge & Robinson, 2017). *Word frequency* refers to how often a word appears relative to the total word count. Figure 14.6 presents a scatter plot of words used in both feedback conditions, where words near the identity line have similar relative frequencies in each condition. This plot shows that most words cluster around the identity line, indicating that most words occur with similar relative frequency in both feedback types. For instance, '*attention*' and '*both*' appeared almost equally in both SA and PP feedback. This observation is supported by the calculation of Pearson's correlation coefficient for word frequencies in both conditions, yielding a high positive correlation of $r(928) = 0.89$.

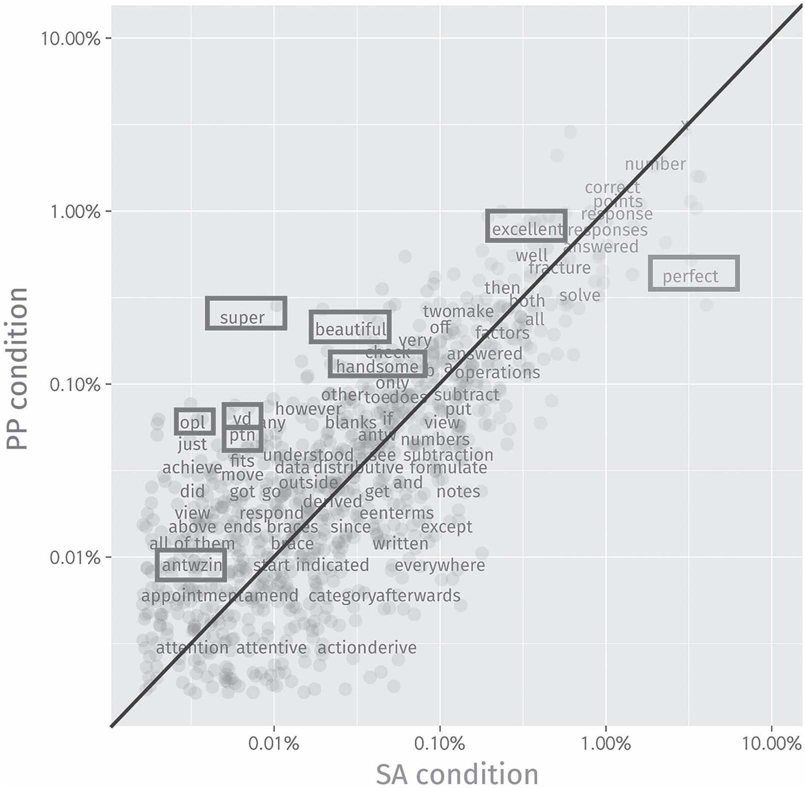


FIGURE 14.6 Comparing the word frequencies of SA and PP feedback.

Words that deviate from the identity line are found more frequently in one feedback type than the other. For example, '*super*' and '*beautiful*' were more common in PP feedback, whereas '*perfect*' was more common in SA feedback. This difference might be due to the default presence of a 'Perfect' button in the SA condition for correct solutions. In contrast, in the PP condition, teachers had to write comments themselves, leading them to use a broader range of positive words. Additionally, the PP feedback showed a higher occurrence of abbreviations like '*opl*' (a Dutch abbreviation for solution), '*vd*' (meaning 'of the'), '*ptn*' (short for 'points'), and '*antw*' (short for 'answer'). The use of these abbreviations is a known coping mechanism teachers employ to manage the workload associated with providing feedback (Price et al., 2010). The re-usability of feedback items in the SA condition seems to discourage frequent use of abbreviations.

For the qualitative analysis, a code was developed to code all feedback reports on students' solutions to the word problem (see Figure 14.7). Two independent raters coded all feedback reports without seeing each other's codes. Four iterations were needed to achieve high interrater reliability, with the final Cohen's kappa coefficients provided in Table 14.2. Feedback was coded student by student, with each rater coding the feedback for 30 out of 60 students.

The codebook distinguished between *categorisable* and *not further categorisable* feedback. Not further categorisable feedback includes erroneous feedback, incomprehensible feedback, or only addressing a solution was perfect, totally wrong, or left blank.

Categorisable feedback is divided into four subcategories:

1. *Concreteness.* Assesses specificity.
   - *General feedback* is applicable to various mistakes.
   - *Concrete feedback* targets a particular mistake.
2. *Focus of the feedback.* Counts the number of deficits and strengths addressed.
3. *Diagnostic activity.* Differentiates between analysis, correction and description.
   - *Analysis.* Interprets the student's (erroneous) reasoning and provides that interpretation as feedback.
   - *Correction.* Points out mistakes and provides the correct solution.
   - *Description.* Identifies deficits without offering a correction.

   Analysis is prioritised over correction, which is prioritised over description.
4. *Quality features of diagnosis.* Includes four aspects that are not mutually exclusive:
   - *Explanation for deficits.* Explains why something is wrong.
   - *Hints for improvement.* Suggests how to improve in future reviews.
   - *Notes missing parts.* Highlights missing elements in the solution.
   - *Points to misconceptions.* Identifies known misconceptions or reasoning errors.

The final results of the analysis are summarised in Table 14.2. The first column displays Cohen's kappa (κ) for the final iteration of establishing the codebook, measuring inter-rater agreement on the final codes in the codebook. Intra-class correlation coefficients are provided for the number of deficits and strengths. The second and third columns show the proportion (in percentages) of feedback reports, in which each code applies in the SA and PP conditions. The mean and standard deviation are reported for the number of deficits and strengths. The fourth column presents p-values from two-sample z-tests comparing the proportions in columns 2 and 3. For the number of deficits and strengths, comparisons were conducted using the Mann–Whitney U test due to the non-normal distribution and independent groups; the U values are noted as a footnote. Finally, the fifth column reports the Spearman's ρ,

TABLE 14.2 Results of coding all feedback reports in both conditions

	κ	*SA* *(n = 913)*	*PP* *(n = 947)*	*p-value*	*ρ*
Categorisable feedback		***78.64%***	***78.04%***	0.631	
Concreteness					
General***	0.82	39.65%	30.52%	<0.001	0.47
Concrete***	0.75	38.99%	47.52%	<0.001	0.51
Focus of the feedback					
Number of deficits**	0.89[1]	1.73 ± 1.28	1.57 ± 1.08	0.003[3]	0.48
Number of strengths*	0.89[1]	0.79 ± 1.05	0.65 ± 0.84	0.038[4]	0.60
Diagnostic activity					
Analysis*	0.88	5.15%	7.60%	0.038	0.22
Correction	0.87	16.21%	16.79%	0.726	0.68
Description	0.84	56.63%	52.80%	0.103	0.58
Quality features					
Explanation for deficits*	0.58	9.42%	12.57%	0.030	0.17
Hints for improvement	1.00	19.72%	23.23%	0.072	0.68
Notes missing parts	0.49	15.55%	14.36%	0.478	0.65
Points to misconceptions	0.82	4.93%	5.07%	0.889	0.31
Not further categorisable feedback		*21.36%*	*21.96%*	0.538	
Erroneous feedback	0.79	4.60%	4.96%	0.711	−0.02
Incomprehensible feedback**	1.00	1.20%	0.11%	0.003	−0.04
Only addresses					
. . . solution is entirely correct*	–[2]	11.17%	14.68%	0.024	0.19
. . . question is left blank*	–[2]	3.40%	1.90%	0.044	−0.20
. . . solution is entirely wrong	–[2]	0.99%	0.32%	0.072	0.52

**p <0.05, **p <0.01, ***p <0.001*

[1] *Intra-class correlation coefficient.*

[2] *Automatically coded.*

[3] *U = 296748 (SA as the reference group).*

[4] *U = 301446 (SA as the reference group).*

Source: Moons, Holvoet, et al. (2024).

reflecting correlations between the frequency of code selection across conditions for each teacher.

Overall, the results in Table 14.2 suggest that SA feedback is less tailored to the student's solution than PP feedback is. The SA reports are almost equally likely to be categorised as general or concrete (39.65% and 38.99%, respectively), while the PP condition produced significantly more concrete feedback (47.52%). However, SA feedback appears to be more detailed, addressing a broader range of deficits and strengths, whereas PP feedback tends to focus more on the primary issues in the solution. A common use of SA, observed in its general application across various deficits and strengths, is as a type of checklist, as illustrated by the following feedback:

- *Choosing the unknown:*
    - *You are confusing the distinction between the number of questions and the points received.*
- *Setting up and solving the equation:*
    - *You did not include the unanswered questions.*
    - *Your equation is simpler than the equation to solve the question, but the solution is right.*

Concerning the diagnostic activity, we see that there are significantly more feedback reports analysing where it went wrong with a student's solution in PP, from which an example is given:

> *Please try again with x being the number of correct answers. Indeed, you know that for 26 questions, he got points. So you express the number of unanswered questions in terms of x. When setting up the equation, you noted 120 instead of 102. You have to take into account the 5 points per correct question.*

SA feedback reports tended to rely more on description and correction as the primary diagnostic activities. Notably, there was a low correlation (0.22) between teachers' analysis use across the two conditions: teachers who engaged in analysis in one condition did not necessarily do so in the other. This suggests that the SA system may discourage teachers from providing feedback that thoroughly analyses students' mistakes. One possible reason is that teachers might intuitively use SA more as a checklist, which can prevent them from fully interpreting the connections between the intermediate steps students took. Indeed, also in well-analysable solutions like the one in Figure 14.7, where the student makes a circular argument using the same information twice, only 5 of the 36 teachers (14%) responded to this fallacy with feedback who analysed the mistake. The other teachers gave descriptive feedback, just noticing simple facts (*'Equation is wrong!')*, or corrective feedback. Of these 5 teachers, only 1 analysed this solution in the SA condition.

The significantly lower number of explanations provided in the SA condition (the only significant difference in quality features) supports this view. PP feedback often targets specific mistakes, with teachers sometimes adding extra explanations. In contrast, SA feedback tends to address all the mistakes in a solution but does so more superficially, offering less focus on the main issues in a student's solution.

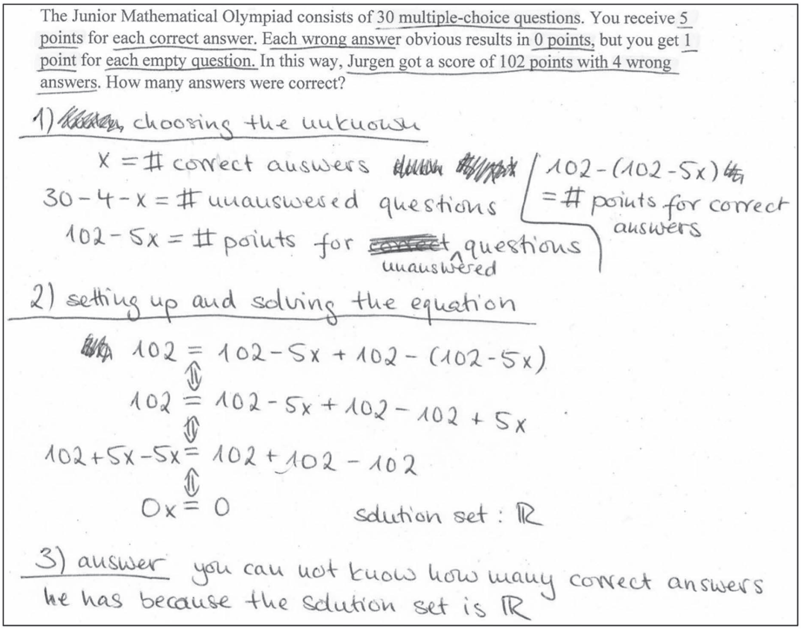


FIGURE 14.7 An 'analysable' student answer.

4.5 Summary of the research results

- Atomic feedback makes feedback significantly more re-usable, proving the usefulness of the guidelines.
- The current SA tool for atomic feedback does not (yet) lead to time savings but to significantly more feedback.
- However, more feedback does not necessarily mean better feedback. Some teachers may be inclined to re-use feedback carelessly, simply describing and correcting students' work rather than analysing underlying misconceptions or misunderstandings. Therefore, it is crucial to continue paying attention to the quality of feedback.

5 Applications of atomic feedback

5.1 Statement banks

A *statement bank* is a collection of pre-written sentences or phrases that can be used to quickly and efficiently compose written feedback (Denton & McIlroy, 2018). Statement banks are often used in educational settings, where teachers might use them to provide standardised comments on student performance. The idea is to have a ready-to-use set of statements that

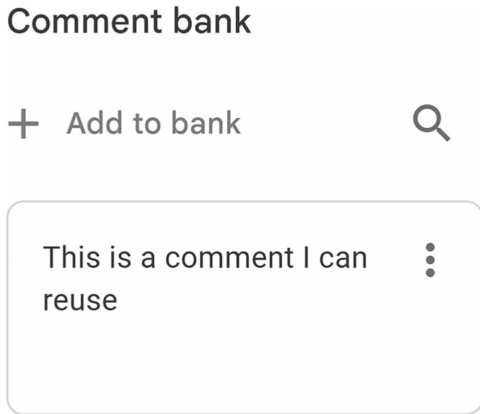


FIGURE 14.8 Comment bank in Google Classroom.

can be customised or directly applied, saving time and ensuring consistency in messaging. Most learning management systems nowadays feature statement banks, like Google Classroom, that allows teachers to maintain a comment bank to store commonly recurring remarks (see Figure 14.8; Google, 2021).

The digital SA tool is an example of a statement bank, and atomic feedback can be viewed as an underlying framework for making the statements as re-usable as possible.

In 2015, Denton and Rowe conducted a study on the impact of feedback derived from statement banks. They found limited evidence of learning from such feedback, which they attributed to the generally low quality of feedback in statement banks. These findings align with our results discussed in Section 4.4.

5.2 Checkbox grading

Checkbox grading is a semi-automated method developed to assess large-scale mathematics exams involving multiple assessors (Moons et al., 2025; Moons, Iannone, et al., 2024). Students write solutions on paper, which are then scanned for correction using the checkbox grading system on a computer. Each task includes a grading scheme with checkboxes that provide (1) *atomic* feedback items and (2) partial points for specific errors. These checkboxes are designed to anticipate common mistakes.

Assessors simply select applicable checkboxes for each student's solution (see Figure 14.9). Dependencies between checkboxes ensure consistency by guiding assessors through a structured grading path. Once grading is

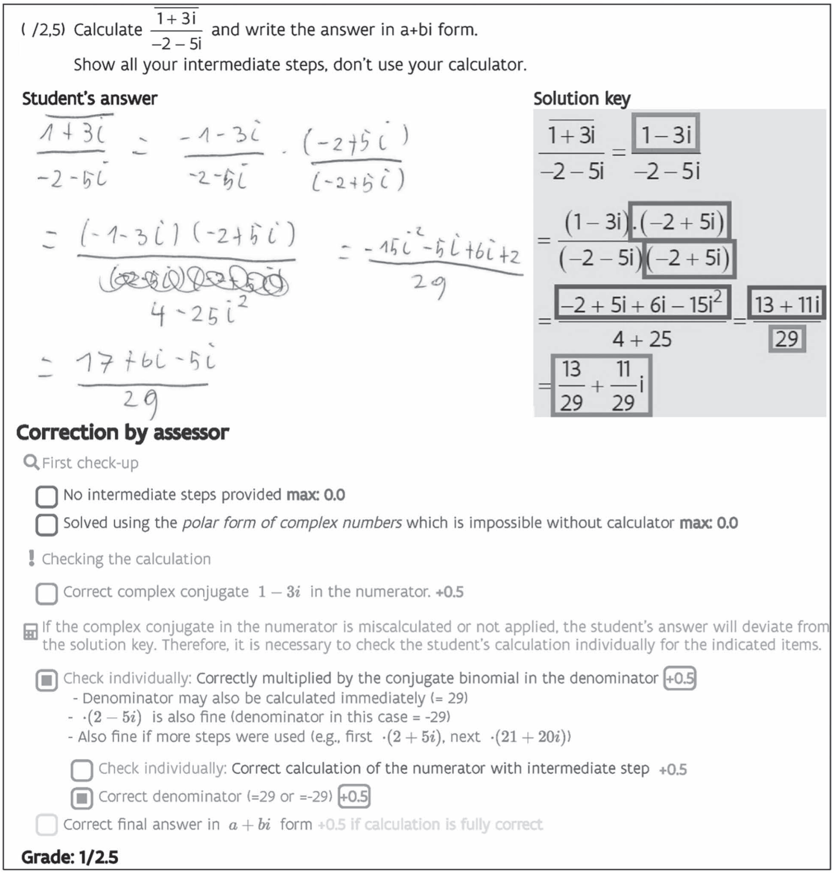


FIGURE 14.9 An example of checkbox grading.

complete, the system generates individual student reports, including grades and feedback from the checkboxes.

Checkboxes can add or subtract points or impose thresholds (e.g. 'No points if this box is ticked,' shown in red in Figure 14.9). They form a hierarchical, point-by-point list that doubles as a series of implicit yes/no questions. Related checkboxes can share colours or clustering to visually link similar steps in a solution (e.g. the indentation in Figure 14.9). This adaptive grading system functions like a flowchart that determines the final grade, ensuring accuracy and minimising personal interpretation.

Research by Moons et al. (2024) demonstrated that checkbox grading provides effective, easily interpretable feedback for high-stakes mathematics exams, regardless of students' level of subject matter knowledge. Assessors found the system highly useful (Moons et al., 2025a), although it requires more time compared to traditional grading, with both methods showing comparable reliability overall. Notably, when strict adherence to grading guidelines is required, checkbox grading significantly enhances inter-rater reliability (Moons & Vandervieren, 2025), particularly through the use of blind grading. In blind grading, the checkboxes only contain feedback; assessors have no insights in how the checkboxes influence grades. Interestingly, despite the increased time investment, assessors perceived the system as more efficient, highlighting a discrepancy between subjective experience and objective time measurements.

Checkbox grading is fundamentally built on the idea of sharing atomic feedback among a group of assessors. Since the checkboxes are shared, an additional criterion for atomicity is introduced: a knowledgeable assessor must be able to unambiguously determine whether a checkbox applies to a student's answer (added to the three guidelines in the definition in Section 2). In essence, each checkbox represents an implicit yes/no question. These four criteria of atomic feedback collectively guide the development of a checkbox grading scheme, with the level of detail in the scheme closely tied to how atomic the feedback is formulated.

6 Next steps: integrating recommender systems with the SA tool for providing atomic feedback

In the experiment described earlier, many teachers acknowledged they sometimes forgot how they had phrased feedback items using the SA tool. As such, they could not find the feedback item they needed, although they knew they had already written a fitting item. This caused them to formulate feedback that had already been given again instead of re-using feedback, confirmed by identifying nearly identical feedback items in their databases by the researcher. The non-intelligent suggestion system was the main reason: It only literally matched what teachers were typing with their items in the database. Improving the suggestion system by incorporating ideas from the extensive literature on recommender systems (Mohanty et al., 2020), is a priority for further research. It is a vital gap to make the semi-automated approach valuable and adopted by teachers. It is also a necessary step before making the software available.

A *recommender system* is a software application or algorithm designed to analyse user preferences, behaviours, and patterns to provide personalised suggestions or recommendations. These systems are commonly used in various online platforms, such as e-commerce websites, streaming services, and social media. Recommender systems aim to enhance user experience by offering relevant content, products, or services tailored to individual tastes and preferences. There are different approaches to building recommender systems, including collaborative filtering, content-based filtering, and hybrid methods, each with its unique way of predicting and delivering recommendations (Li et al., 2024). However, we experience some limitations with our integration: first, we experience a so-called cold start; no items are available for re-use for a new task. Moreover, we can only rely on content-based filtering as we cannot suggest feedback from other teachers.

Currently, we are experimenting with feeding our recommending system with the following information:

- *Item popularity.* First, suggest more popular items by simply counting the number of times a feedback item was re-used.
- *Item distance.* Calculate the average distance of a feedback item to already-selected items. The distance is measured by how many times items co-occurred previously. Closer items are suggested first.
- *Error location.* We now allow teachers to indicate where an error has occurred in a handwritten solution, which provides helpful information about which feedback items are appropriate. However, this is not perfect, as some students have different ways of writing down their solutions.

The first results indicate that a recommender system can make the re-use of feedback items much more straightforward. However, while recommender systems offer a promising starting point, addressing teachers' difficulties in recalling or locating previously created feedback could benefit from alternative approaches. Techniques such as semantic search, improved pattern matching, and generative AI (GenAI) could enhance the system's ability to retrieve (or even create) relevant feedback items more effectively. These approaches could complement traditional recommender systems by enabling more flexible and context-aware feedback suggestions, particularly when exact matches are unavailable.

7 Summary and concluding remarks

In this chapter, we have examined integrating digital technologies into mathematics assessment, focusing on enhancing the efficiency of feedback delivery. One key characteristic of mathematical assessment is that incorrect answers often exhibit structural error patterns across students, leading to repeated feedback and similar grading. This leads us to the concept of atomic feedback, which involves a set of format requirements for mathematical feedback. Our findings demonstrate that atomic feedback increases the re-usability of feedback.

Interestingly, while re-using feedback can lead to more comprehensive feedback, it does not necessarily result in time savings for teachers or higher-quality feedback. Feedback generated through our semi-automated tool was often more elaborate but less tailored to individual student solutions. In contrast, traditional pen-and-paper feedback was shorter but more focused and concrete, directly addressing key issues. This suggests that using semi-automated feedback tools can sometimes shift teachers' focus away from effective diagnostic activities. As such, ongoing attention to the quality and relevance of feedback remains crucial.

We presented two primary applications of atomic feedback: statement banks and checkbox grading. Statement banks naturally lend themselves to creating atomic feedback, while checkbox grading integrates atomic feedback directly into the assessment process. In this system, assessors tick checkboxes corresponding to specific aspects of a student's solution, with each checkbox representing an implicit yes/no question tied to common errors or correct solution steps. This structured approach ensures consistency across assessors by minimising personal interpretation and guiding assessors through a clear grading path. The system then calculates the grade and generates detailed feedback reports, offering students clear insights into their errors and the grading process. Although this approach requires more time than traditional grading, assessors found it subjectively more efficient, likely due to the clarity and structure it provides. Moreover, while objective inter-rater reliability remains comparable to traditional methods, the system fosters greater transparency, which both assessors and students highly value. Notably, even lower-performing students were able to interpret checkbox grading feedback with ease.

Looking ahead, the future of feedback re-use in mathematics education is promising and necessary. As structural error patterns ensure that students will continue to make similar mistakes, this inevitability can be transformed into an opportunity. While this chapter introduced a first attempt using a self-developed Moodle plugin, improving the suggestion system for feedback items remains key.

Reference list

Bokhove, C., & Drijvers, P. (2010). Digital tools for algebra education: Criteria and evaluation. *International Journal of Computers for Mathematical Learning*, *15*(1), 45–62. https://doi.org/10.1007/s10758-010-9162-x

Busch, J., Barzel, B., & Leuders, T. (2015a). Die Entwicklung eines Instruments zur kategorialen Beurteilung der Entwicklung diagnostischer Kompetenzen von Lehrkräften im Bereich Funktionen. *Journal für Mathematik-Didaktik*, *36*(2), 315–338. https://doi.org/10.1007/s13138-015-0079-8

Busch, J., Barzel, B., & Leuders, T. (2015b). Promoting secondary teachers' diagnostic competence with respect to functions: Development of a scalable unit in continuous professional development. *ZDM*, *47*(1), 53–64. https://doi.org/10.1007/s11858-014-0647-2

Denton, P., & McIlroy, D. (2018). Response of students to statement bank feedback: The impact of assessment literacy on performances in summative tasks. *Assessment & Evaluation in Higher Education*, *43*(2), 197–206. https://doi.org/10.1080/02602938.2017.1324017

Denton, P., & Rowe, P. (2015). Using statement banks to return online feedback: Limitations of the transmission approach in a credit-bearing assessment. *Assessment & Evaluation in Higher Education*, *40*(8), 1095–1103. https://doi.org/10.1080/02602938.2014.970124

Eurydice. (2021). *Teachers in Europe: Careers, development and well being [Eurydice report]*. Publications Office of the European Union. https://doi.org/10.2797/997402

Flemish Education Inspectorate. (2023). *Onderwijsspiegel 2023. Jaarlijks rapport van de onderwijsinspectie. (Dutch) [Year report 2023 of the Flemish education inspectorate]*. Vlaamse Overheid. www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1), 9. https://doi.org/10.1186/s40594-021-00323-x

Gibson, S., Oliver, L., & Dennison, M. (2015). *Workload challenge: Analysis of teacher consultation responses* (Vol. 355). Department for Education London.

Google. (2021). *Give feedback on assignments in Google classroom*. https://support.google.com/edu/classroom/answer/9093530?hl=en#zippy=%2Cadd-and-save-comments-to-use-later

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. https://doi.org/10.3102/003465430298487

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Knight, N. (2003). Teacher feedback to students in numeracy lessons: Are students getting good value? *Set: Research Information for Teachers*, *3*, 40–45. https://doi.org/10.18296/set.0704

Lemmo, A. (2023). Tasks in paper and digital environments: An exploratory qualitative study. *International Journal of Mathematical Education in Science and Technology*, 1–21. https://doi.org/10.1080/0020739X.2023.2268613

Li, D., Lian, J., Zhang, L., Ren, K., Lu, T., Wu, T., & Xie, X. (2024). *Recommender systems: Frontiers and practices*. Springer Nature. https://doi.org/10.1007/978-981-99-8964-5

Lim, K. H., Benbasat, I., & Todd, P. A. (1996). An experimental investigation of the interactive effects of interface style, instructions, and task familiarity on user performance. *ACM Transactions on Computer-Human Interaction*, *3*(1), 1–37. https://doi.org/10.1145/226159.226160

Mohanty, S. N., Chatterjee, J. M., Jain, S., Elngar, A. A., & Gupta, P. (Eds.). (2020). *Recommender system with machine learning and artificial intelligence: Practical tools and applications in medical, agricultural and other industries*. John Wiley & Sons, Inc.

Moons, F. (2023). *Semi-automated assessment of handwritten mathematics tasks atomic, reusable feedback for assessing student tests by teachers and exams by a group of assessors*. Doctoral dissertation, University of Antwerp. https://hdl.handle.net/10067/1980770151162165141.

Moons, F., Holvoet, A., Klingbeil, K., & Vandervieren, E. (2024). Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques. *British Journal of Educational Technology*, *55*(5), 2257–2277. https://doi.org/10.1111/bjet.13447

Moons, F., Iannone, P., & Vandervieren, E. (2024). Checkbox grading of handwritten mathematics exams with multiple assessors: How do students react to the resulting atomic feedback? A mixed-method study. *ZDM – Mathematics Education*. https://doi.org/10.1007/s11858-024-01550-6

Moons, F., & Vandervieren, E. (2025). Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa. *Behavior Research Methods, 57*, 287. https://doi.org/10.3758/s13428-025-02746-8

Moons, F., Vandervieren, E., & Colpaert, J. (2022). Atomic, reusable feedback: A semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers. *Computers and Education Open*, *3*, 100086. https://doi.org/10.1016/j.caeo.2022.100086

Moons, F., Vandervieren, E., & Colpaert, J. (2025). Checkbox grading of large-scale mathematics exams with multiple assessors: Field study on assessors' inter-rater reliability, time investment and usage experience. *Studies in Educational Evaluation*, *85*, 101443. https://doi.org/10.1016/j.stueduc.2024.101443

Movshovitz-Hadar, N., Zaslavsky, O., & Inbar, S. (1987). An empirical classification model for errors in high school mathematics. *Journal for Research in Mathematics Education*, *18*(1), 3–14. https://doi.org/10.2307/749532

Olsher, S., Chazan, D., Drijvers, P., Sangwin, C., & Yerushalmy, M. (2024). Digital assessment and the "machine". In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 1175–1201). Springer International Publishing. https://doi.org/10.1007/978-3-031-45667-1_44

Pelkola, T., Rasila, A., & Sangwin, C. (2018). Investigating bloom's learning for mastery in mathematics with online assessment. *Informatics in Education*, *17*(2), 363–380. https://doi.org/10.15388/infedu.2018.19

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Schnepper, L. C., & McCoy, L. P. (2014). Analysis of misconceptions in high school mathematics. *Networks: An Online Journal for Teacher Research*, *15*(1).

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1st ed.). O'Reilly.

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*(3), 335–348. https://doi.org/10.1007/s10649-006-9078-5

15

DEVELOPING MATHEMATICAL COMPETENCIES TO ASSESS CHATGPT-GENERATED OUTPUTS

Three forms of mediations

Marianne Thomsen, Morten Misfeldt, and Uffe Thomas Jankvist

1 Introduction

The rise of generative artificial intelligence (GAI) is revolutionising the educational system, presenting both opportunities and challenges. These new tools allow students to develop content and solve tasks with minimal engagement, posing significant concerns to teachers and educators. However, the potential for personalised learning environments and student empowerment makes it impossible to ignore AI's potential positive role in mathematics education. As argued elsewhere (Geraniou et al., accepted), this debate mirrors the earlier inclusion of digital tools in mathematics teaching, highlighting valuable similarities and differences as we navigate AI integration in education. The use of CAS in mathematics teaching has taught us several valuable lessons that are applicable as we integrate artificial intelligence into education. Researchers, educators, teachers, and students cannot omit discussing what processes that students can or cannot outsource at different stages of their mathematical work. Hence, we need to be very aware to what extent tools are used for learning mathematics. And if the tools are merely handling mathematical processes so that the students do not need to worry about them.

From an assessment perspective, the chapter explores how GAI – we focus particularly on ChatGPT – can help students develop their mathematical reasoning, thinking, and communication competencies. We aim to address how to teach students to use ChatGPT to investigate mathematical texts and how to focus on formative assessment situations in such work. We focus on primary historical sources to emphasise a different approach to mathematical activity from the typical task delivery paradigm in mathematical education.

DOI: 10.4324/9781003533764-15

From a theoretical standpoint, the chapter focuses on how the terms *epistemic*, *pragmatic*, and not least *justificational mediations* can be important lenses to qualify formative assessment, as well as on how to support students' possibilities to develop their reasoning, communication, and thinking competencies. We view this related to learning setups where the potential role of ChatGPT is to enhance students' ability to critically analyse and articulate their understanding of mathematical texts, concepts, and arguments. Related hereto, we also use the terms *techno-authoritarian proof schemes*. The chapter highlights the importance of teaching students to navigate between beneficial ChatGPT–mediated learning and the pitfalls of uncritical acceptance of ChatGPT outputs and how to assess the students' competency in doing so.

In the following, we first describe the framework of mathematical competencies we rely on. Next, we present a section on mediations and one on assessment. Then we have a section where we shortly describe the primary historical source we use together with ChatGPT. We focus on Euclid's five postulates, Book I, and Euclid's Proposition 6, Book IV: *To inscribe a square in a given circle*. We then combine the theoretical terms in an analytical frame. We use this frame to analyse some excerpt from one of our own chats with ChatGPT-4.[1] This has the purpose of exemplifying how the analytical frame can be used in formative assessment situations, which include the use of ChatGPT and primary historical sources. We end with a few concluding remarks.

2 Mathematical competencies

The Danish KOM framework defines eight mathematical competencies (Niss & Højgaard, 2011, 2019). A *mathematical competency* is defined as 'someone's insightful readiness to act appropriately in response to *a specific sort* of mathematical *challenge* in given situations' (Niss & Højgaard, 2019, p. 14). The eight competencies are mathematical thinking, problem handling, modelling, reasoning, representation, symbols and formalism, communication and aids, and tools competency. In this chapter, we focus mainly on the three competencies of mathematical reasoning, mathematical thinking, and mathematical communication. We describe each of these in what follows and then relate these competencies to the use of ChatGPT. First, however, we point the reader's attention to a key feature of the KOM framework's competencies, namely, their dual nature. Each competency has a 'receptive facet' and a 'constructive facet' (Niss & Højgaard, 2019). The receptive facet involves an individual's ability to relate to and navigate existing considerations and processes introduced by others. For example, this includes following and assessing a mathematical proof or a proposed solution. The constructive facet focuses on the individual's ability to independently utilise the competency for constructive purposes. This includes, for example, devising a mathematical proof, posing mathematical questions, or simplifying a complex expression.

The core of the *mathematical reasoning competency* is the ability to analyse or produce arguments (i.e. chains of statements linked by inferences) that justify mathematical claims. This involves both constructing justifications and critically assessing existing or proposed ones. It encompasses various forms of justification, from providing examples and counterexamples to using heuristics, local deduction, and rigorous proofs based on logical axioms. Niss and Højgaard (2019) state: ‘It is important to stress that the kinds of claims at issue in this competency are not confined to “theorems” or “formulae” but comprise all sorts of conclusions obtained by mathematical methods and inferences, including solutions to problems’ (p. 16).

The *mathematical thinking competency* involves posing and relating to the types of questions characteristic of mathematics and understanding the nature of expected answers. It includes recognising the varying scope of mathematical concepts in different contexts, distinguishing between different types of mathematical statements (such as definitions, if–then claims, universal and existence claims, singular case statements, and conjectures), and understanding the roles of logical connectives and quantifiers in these statements. Additionally, it encompasses proposing abstractions of concepts and theories and generalisations of claims as part of mathematical activity.

Oftentimes, aspects of the reasoning and thinking competencies are confused with one another. However, there is a clear division of labour between the two, which means they do not actually represent the same thing. As mentioned earlier, the reasoning competency focuses on the justification of mathematical claims. In contrast, the thinking competency focuses on considering what kinds of questions are asked in mathematics and what kinds of answers can be expected, without concerning itself with the production or justification of answers to specific questions. You might say that you use your thinking competency when you figure out what would be good questions to ask ChatGPT about related to the mathematical ideas in primary historical sources.

Mathematical communication competency involves the ability to engage in different kinds of mathematical communication, ‘in different genres, styles, and registers, and at different levels of conceptual, theoretical and technical precision, either as an interpreter of others’ communication or as an active, constructive communicator’ (Niss & Højgaard, 2019, p. 18). The communication competency is also about involving the awareness of senders and recipients in the communication (Niss & Jensen, 2002, p. 63). While students communicate with ChatGPT, they communicate with the computer feedback on their own questions. They really must be aware of how they prompt and how they analyse and interpret the feedback from ChatGPT. Therefore, we see the communication competency as an important competency to assess in light of students’ use of ChatGPT.

The reasoning competency is about understanding peoples’ or computers’ justification and of the production of mathematical arguments, the thinking competency is about understanding and creating relevant questions, while the communication competency is about individuals' ability to engage in mathematical communications. In this chapter, we mainly focus on the reasoning competency, but we find the two other mathematical competencies essential, while working with assessments situations where both ChatGPT and primary historical sources are in focus.

3 Mediations

Aligned with Tamborg (2021), we define the term 'mediation', with reference to Rabardel and Bourmaud's (2003) activity theory (Nardi, 1996), as 'the fact that the subject's use of the artifact influences the subject or makes something possible for the subject to do' (Tamborg, 2021, p. 1,062). Misfeldt and Jankvist (2018) differentiate between three types of mediation when working with digital technology: *epistemic*, *pragmatic*, and *justificational mediation*. Their work builds on the foundations laid by Rabardel and Bourmaud (2003), as well as the distinction between epistemic and pragmatic value (e.g. Artigue, 2002) and epistemic and pragmatic functions (e.g. Trouche, 2005). This framework originates from the instrumental approach, which is commonly applied in mathematics education to examine how students use technology to learn mathematics. It considers the appropriation of digital tools for solving mathematical problems, viewing computational artefacts as mediators between the user and their goal. The approach suggests that a student's goal-directed activity is shaped by using tools – a process often referred to as instrumentation – while simultaneously, the goal-directed activity reshapes the tool – a process known as instrumentalisation.

In the context of students' interaction with technology, a distinction is made between epistemic and pragmatic mediations. *Epistemic mediations* concern goals internal to the user, such as influencing their understanding, perspective, or knowledge of a subject. For example, Rabardel and Bourmaud (2003) discuss the use of a microscope, while Lagrange (2005) highlights experimental uses of computers. *Pragmatic mediations*, on the other hand, relate to goals external to the user, such as effecting a change in the world. Rabardel and Bourmaud (2003) use the example of a hammer, and Lagrange (2005) refers to the mathematical technique of 'pushing buttons'. Notably, computer algebra systems (CAS) serve both pragmatic and epistemic purposes (Artigue, 2002; Trouche, 2005).

Misfeldt and Jankvist (2018) combine the aforementioned distinction between mediations with Hanna's (1990) distinction between proofs that (only) prove and proofs that prove and explain, as well as Harel and Sowder's (2007) description of (conviction) proof schemes, that is, what constitutes ascertaining and persuading for a person or community. According to Harel and Sowder (2007), a *deductive proof scheme* is linked to being convinced by logical deduction; an *external convicting proof scheme* is linked to being convinced by an authority, for example, a textbook or a teacher, whereas an *empirical proof scheme* is linked to use of empirical examples for justifying mathematical statements. Jankvist and Misfeldt (2019) explain:

> *Epistemic mediations* are connected to proofs that explain (Hanna, 1990), as well as to deductive proof schemes (Harel & Sowder, 2007). *Justificational mediations* are related to proofs that only proves, i.e. without explaining. Furthermore, such mediations are connected to external conviction proof schemes. If statements are true because CAS says so, CAS mediates a justificational process. *Pragmatic mediations* may be connected to one or more of the different proof schemes, including the empirical proof scheme, by providing necessary but laborious calculations and manipulations required for a certain argument.
>
> *(p. 249)*

This description of justificational mediations is, to some extent, connected to what Misfeldt and Jankvist (2018) call *techno-authoritarian proof schemes*, which is when the computer output becomes an authority. Misfeldt and Jankvist (2018) used these three types of mediation to analyse the didactical effects of the use of CAS-assisted proof in Danish upper secondary school mathematics textbooks. The three types of mediation can also be used related to DGS (e.g. Thomsen & Jankvist, 2020). When it comes to using ChatGPT, we argue that justificational mediation might play another role than in, for example, CAS and dynamic geometric software. Here, justificational mediation includes a more linguistic dialogical aspect. On the one hand, this has the risk of making the computer output even more convincing, that is, more techno-authoritarian, because ChatGPT also can formulate (chains of) mathematical arguments linguistically. At first glance, it looks like it is a proof that explains, but it might not be the case for the user and then it is still a justificational mediation. In other words, we see a risk that justificational mediation will be the most occurring mediation. On the other hand, it may also be easier for the user to question and challenge computer-generated arguments. Hence, we deem it important to pay special attention to justificational mediation, while students work with ChatGPT, in order not to encourage the promotion of techno-authoritarian proof schemes. It becomes essential to focus on how to support students' creativity and to critically question, by prompting a way, where they also develop an awareness of when justificational mediation supports an epistemic mediation or a pragmatic mediation, and whether the outcome is mathematically reasonable or not. Misfeldt and Jankvist (2018) have phrased four questions related to the use of CAS to structure the discussion of CAS use in relation to proofs and proving (Misfeldt & Jankvist, 2018, p. 283). Jankvist et al. (2019) rely on these and say:

> These four aspects of justificational mediations are:
>
> 1. Does the CAS use establish truth? This is the core function of a justificational mediation. To what extent does the CAS output act as warrant in an argument?
> 2. Does the CAS use allow interaction and experimentation? This highlights the degree to which students can change parameters, explore phenomena etc., and therefore to what extent the students still have agency when working with CAS in relation to proofs.
> 3. Is the argumentation inductive, deductive or authoritarian? What type of proof scheme is in play and what type of warrant do CAS provide.
> 4. Does the argument highlight important aspect of the proof or the mathematical relationships? Both the second and the fourth aspects are to some extent related to epistemic mediations.
>
> *(Jankvist et al., 2019, p. 326)*

Even though there is a difference between using CAS, DGS, or ChatGPT, we still find these four questions relevant when it comes to the use of ChatGPT in assessment situations. This is because, using ChatGPT, is about assessing both the students' own questions and their ability to reflect on ChatGPT's answers in a critical way so justificational mediation can relate to epistemic meditation. Furthermore, the students must look critically on their way to building up a dialogue with ChatGPT, on how ChatGPT's answer influences their next question and how this corelates with their plan of using ChatGPT to solve a task. Here we find it important to be aware of the three competencies – reasoning, thinking, and communication – to support the students' possibilities to develop a constructive and critical view of their own use of ChatGPT.

4 Assessment

We focus on how assessment can support students' possibilities to develop their mathematical reasoning, thinking, and communication competency while they work with ChatGPT and primary historical sources.

We particularly concentrate on formative assessment, where the purpose is to improve the next step in the students' learning by assessing what they know (Taras, 2005); in other words, this is an assessment 'helping for learning' (Harlen, 2005). In the context of CAS, Jankvist et al. (2021) write about different kinds of assessments' situations:

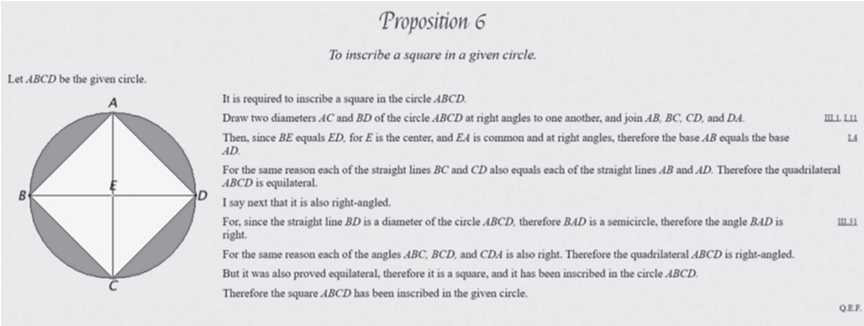
> Answering in what ways CAS augment and change the assessment situation, we can say that our examples show that (1) CAS challenge existing

> assessment practices by moving them in a pragmatic direction; (2) CAS allow for new types of assessment situations to be implemented because of the increased computational power that CAS provide; finally, (3) CAS can also be used as a part of the teachers' setup of the assessment situation. By having computer algebra at hand in automated and real-time student response systems, CAS can enrich the immediate feedback that students receive.
>
> *(pp. 115–116)*

If we imagine how ChatGPT augments and changes the assessment situation, where the students can use CAS and DGS, and look at number 1 earlier, we find it likely that assessment practices can risk moving in a justificational direction. In an assessment situation it might be difficult to distinguish between a justificational mediation and a pragmatic or epistemic mediation because of ChatGPT's power of linguistics persuasion. Even though the students use ChatGPT-generated answers to explain or formulate, for example, steps in mathematical reasoning, it may seem to be an epistemic mediation while it is, in fact, a justificational mediation that can be connected to a techno-authoritarian proof scheme. In assessment situations, ChatGPT may also have the potential to move in a more epistemic direction, which would require that the students have a creative and critical approach to their use of ChatGPT in such situations. Surely, it is also possible to ask critical questions to the output, but if one is not able to see through the arguments and explanations provided, there is a risk that justificational mediation turns into a techno-authoritarian proof scheme.

If we look at Jankvist et al.'s (2021) number 2 in the preceding quote, ChatGPT will probably also allow for the implementation of new types of assessment situations. For example, it would be obvious to let the students engage in a mathematical 'dialogue' with ChatGPT. The teacher could also let the students use ChatGPT as part of the assessment situation, where their 'dialogue' with ChatGPT can enrich the feedback they receive while they also work with other digital tools, for example, CAS or DGS, as well as with their own drawings and formulation of arguments. This may be a way of supporting the second and the fourth aspects listed by Misfeldt and Jankvist (2018) and Jankvist et al. (2019) in the previous section, that is, to support that justificational mediations can be related to epistemic mediations. We also find that number 3 in the earlier quote from Jankvist et al. (2021) will be a possible consequence of using ChatGPT in assessment situations.

To some extent, one might find that Jankvist et al.'s (2021) number two is implicit in this chapter, because we find using ChatGPT in formative assessment situation is new because of the linguistic power of this tool. We will try to give an analytical frame which hopefully can support not ending in the number 1 assessment type presented earlier and in our empirical example we will try to focus on the number 3 type of assessment described earlier.


Proposition 6

To inscribe a square in a given circle.

Let *ABCD* be the given circle.

It is required to inscribe a square in the circle *ABCD*.

Draw two diameters *AC* and *BD* of the circle *ABCD* at right angles to one another, and join *AB*, *BC*, *CD*, and *DA*. III.1, I.11

Then, since *BE* equals *ED*, for *E* is the center, and *EA* is common and at right angles, therefore the base *AB* equals the base *AD*. I.4

For the same reason each of the straight lines *BC* and *CD* also equals each of the straight lines *AB* and *AD*. Therefore the quadrilateral *ABCD* is equilateral.

I say next that it is also right-angled.

For, since the straight line *BD* is a diameter of the circle *ABCD*, therefore *BAD* is a semicircle, therefore the angle *BAD* is right. III.31

For the same reason each of the angles *ABC*, *BCD*, and *CDA* is also right. Therefore the quadrilateral *ABCD* is right-angled.

But it was also proved equilateral, therefore it is a square, and it has been inscribed in the circle *ABCD*.

Therefore the square *ABCD* has been inscribed in the given circle.

Q.E.F.

FIGURE 15.1 An example of an English translation of Euclid's *Elements*, Book IV, Proposition 6 (Joyce, 1996b).

5 Primary historical sources and ChatGPT

In our choice of a primary historical source, we are inspired by the work of Thomsen (2022), who used excerpts from Euclid's *Elements* in a setting of using DGS (GeoGebra) with a class of sixth and seventh grade students (age 11–14 years). In particular, the students worked with Euclid's five postulates from Book I and Proposition 6, concerning the inscription of a square in a given circle, in Book IV (see Figure 15.1).

The five postulates read:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and radius.
4. That all right angles equal one another.
5. That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

(Joyce, 1996a)

The aim was to support the students in developing their reasoning competency, while they worked with the intersection with the primary historical source and GeoGebra. Besides reasoning, focus was also on the thinking competency, since the task was about understanding Euclidian geometry and understanding what the propositions and the reasoning herein were built upon (see, for example, Thomsen & Jankvist, 2021). We will not go further into the empirical examples related hereto but make use of the same primary historical source and take this as an outset for how our analytical frame can be used to assess students' work with ChatGPT and primary historical sources. Using a primary historical source can give the students the possibility to reflect on ChatGPT's answer related to the argumentation and the main ideas in the source. One might even say that the students both have an outset to plan their prompting with outset in the source, and at the same time they can compare ChatGPT's feedback with the text in the source. Here the communication competency is also in play, because the students can focus on both their communication with ChatGPT and how they read.

6 The analytical frame

When trying to understand how to assess and support students' mathematical work with primary historical sources and ChatGPT – where the aim is to support their possibilities for developing their reasoning, thinking, and communication competencies – we suggest that teachers navigate and have a focused glance on respectively epistemic, pragmatic, and justificational mediations. *Epistemic mediation* involves students gaining knowledge and understanding through interaction with ChatGPT. *Pragmatic mediation* focuses on the practical application of ChatGPT tools to solve problems and achieve specific tasks. Justificational mediation, however, requires students to relate to the ChatGPT output so they not just take the computer output as a correct answer or explanation without reflecting on it. If they do not become aware of this, they risk developing mainly techno-authoritarian proof schemes, which is when students merely take ChatGPT's answer as a right answer because it is computer-generated. Teachers must find a balance between providing necessary guidance and fostering an environment that encourages independent critical inquiry. We find that one setup for such mathematical assessment situations can be letting the students use ChatGPT to understand the text in the primary historical source and to see the content of the text from different perspectives. In this situation, the teacher can both assess the students' prompting of ChatGPT and to what extent the students have a reflective and critical approach to the provided answers. The three types of mediation can be a valuable tool for the teacher, both to assess and to plan new steps or situations which support the student in getting a deeper understanding and possibilities of using this in a constructive way. We have chosen excerpts from Euclid's *Elements* to assess students' potential work with these and ChatGPT, and we find that it will require students to use their reasoning, thinking, and communication competency in their prompting of ChatGPT and the analysis of the outcome. To address this, we propose an analytical assessment framework that incorporates the three types of mediation and the three mathematical competencies. This framework aims to provide a structured yet flexible method to assess and support students' work with the primary historical sources and ChatGPT. The situation is depicted in Figure 15.2. In this chapter we focus on the justificational mediation and how it can switch between

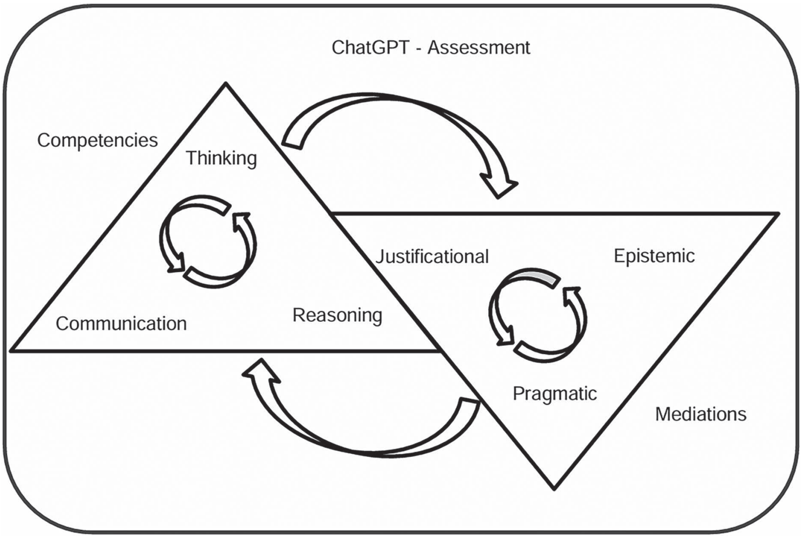


FIGURE 15.2 Composition of the theoretical terms we use in the analysis.

supporting an epistemic or a pragmatic mediation. The arrows in Figure 15.2 show that it is dynamic how these different theoretical distinctions influence one another.

In this chapter, reasoning competency and justificational mediation are in the foreground, shown in the model by the 'corners' that are adjacent to each other between the two triangles. As explained previously, we find that these are related to the two other types of mediations and mathematical competencies, respectively, when it comes to an assessment situation of students' work with ChatGPT and our choice of a primary historical source. When teachers assess students' work with ChatGPT, they must both be aware of the students' use of their mathematical competencies and which mediations seem to be in play in the students' interaction with ChatGPT. While using formative assessment, the teachers must also be aware of what mathematical settings may next support students in their learning process. We find that the analytical frame can be put into play here, so the formative assessment situation supports justificational mediation tipping towards an epistemic mediation to reach the goal. This is mainly to assess the students' reasoning competency and to support their possibilities to further develop it – for some students, it might be beneficial to go through a pragmatic mediation to reach this point. The students must also use their thinking and communication competencies when working with ChatGPT and a primary historical source. In this chapter, we do not focus on assessment situations that mainly have a pragmatic nature, the ones Jankvist et al. characterise as number 1. If that was the case, or if we had another mathematical goal for the assessments, we might look at, for example, the pragmatic mediation and the communication competency and thus rearrange the placement of these in two triangles in Figure 15.2.

7 The use of the analytical frame

In the following, we analyse some selected excerpts of our own chats with ChatGPT-4[2] about understanding the content of Euclid's five postulates, Book I, and Proposition 6, Book IV, and the relation between these. Hence, it is not students' work with ChatGPT. Yet it gives us an impression of how such an assessment situation could play out. In this case, the reasoning competency is in focus, and the thinking and communication competencies are addressed when relevant. We prompt ChatGPT-4, and each question is answered in a rather-long explanation by ChatGPT. If students merely take ChatGPT's answers without understanding them, one might characterise this as a justificational mediation, which eventually may turn into a techno-authoritarian proof scheme. If the students understand ChatGPT's explanations, we can characterise it as an epistemic mediation, but here the teacher also must take other signs on the students' learning processes into play to make an assessment. Of course, the teacher can use the students' prompting of ChatGPT, which to some extent can be a kind of assessment. Still, we think it is needed that the teacher in such a mathematical situation only see ChatGPT as a part of the situation. Hence, we imagine this as the kind of assessment situation that Jankvist et al. (2021) characterise as number 3. Later we will attempt to give some examples of how teachers can orchestrate such situations related to ChatGPT answers, to support students' further learning process.

In our chat with ChatGPT-4, we first asked questions about the five postulates. Here the thinking competency is in the foreground, because this is about understanding the content of postulates that are behind the reasoning according to Euclid's Proposition 6, Book IV (Thomsen & Jankvist, 2021). We asked ChatGPT-4, 'Tell me about Euclid's five postulates'. ChatGPT-4 presented the five postulates and explained them step by step. If the students do not reflect on this answer, it can be characterised as justificational mediation, which may then turn into a techno-authoritarian proof scheme. To avoid such a situation, the teacher can ask the students to talk about the difficult words and language in Euclid's five postulates and maybe compare these with the ChatGPT answer. This approach could be a deliberate way to involve the communication competency. The teacher might also let the students use GeoGebra to construct a visualisation of the five postulates (inspired by Thomsen, 2022). This could support an empirical, and maybe also an epistemic, mediation. In addition, this could make up an assessment situation, which ChatGPT is a part of, and the students then have the possibilities to explore how to draw the postulates in GeoGebra based on their reading of ChatGPT's explanation and the students' own reading of Euclid's five postulates.

Let's break down the proof of Proposition 6 from Euclid's "Elements" and see how it uses the five postulates.

Proposition 6:

To inscribe a square in a given circle.

1. **Given:**
   - A circle ABCD.
2. **Required:**
   - To inscribe a square in the circle ABCD.

Steps of the Proof:

1. **Drawing Diameters (Using Postulates 1 and 3):**
   - Draw two diameters AC and BD of the circle ABCD at right angles to one another.
     - Postulate 1: A straight line can be drawn between any two points (here, the ends of the diameters).
     - Postulate 3: A circle can be drawn with any center and any radius (to identify the circle).
2. **Joining Points (Using Postulate 1):**
   - Join AB, BC, CD, and DA to form a quadrilateral.
     - Postulate 1: A straight line can be drawn between any two points.

FIGURE 15.3 A ChatGPT-4 explanation of how the proof uses the five postulates. See the full chat here: https://chatgpt.com/share/55f7893e-13fb-49b3-81aa-c694966ea216.

We also asked ChatGPT-4 about Euclid's Proposition 6, Book IV, using the five postulates. We prompted ChatGPT-4: 'Help me understand how this proof uses the five postulates [and] "Proposition 6: To inscribe a square in a given circle . . ."'. In the prompt we have pasted the text of Proposition 6, Book IV (see Figure 15.1) in after the query. ChatGPT answered:

One might say that the prompt shows an understanding of the relation between the five postulates, Book I and Proposition 6, Book IV. We interpret the part of the answer which says 'given' and 'required' as justificational mediation supporting an epistemic mediation because it simplifies and explains the title of Proposition 6, Book IV. Then steps 1 and 2 of the proof can also be characterised as supporting an epistemic mediation since it relates the postulates to the proof. One might also say that this way of explaining supports an understanding of the generality in the proof, because it explains by using the postulates, for example, 'A straight line can be drawn between any two points (here, the ends of the diameters)'. We can characterise this as another way of communication other than the students' reading of the source, which does not explicitly mention the postulates. Hence, such interaction has the potential to support students in using and developing their reasoning, thinking, and communication competencies – at least the 'receptive facet'. Here the teacher could also assess and support the students in using the 'constructive facets' of the competencies, by asking them to unfold why the diameters must be perpendicular if it is an equilateral square. What would happen if the diameters were not perpendicular? This might also have the potential to put the students in situations where epistemic mediation is in the foreground. If the teacher let the students use GeoGebra to explore what happens in that case, it might begin with pragmatic mediation, which hopefully could turn into an epistemic mediation.

We also prompted ChatGPT-4 with: 'Can you help me write the argument as a clear proof?' In that case, an excerpt of the answer looked as follows:

3. **Prove the Quadrilateral is Equilateral:**
   - Consider the triangles ABE and ADE:
     - $BE = ED$ because they are radii of the circle.
     - AE is common to both triangles.
     - $\angle BEA = \angle DEA = 90^\circ$ because BD is a diameter and they intersect AC at right angles.
     - By the Side-Angle-Side (SAS) congruence theorem (I.4) $\triangle ABE \cong \triangle ADE$.
     - Hence, $AB = AD$.
   - Similarly, we can show:
     - $\triangle BEC \cong \triangle DEC$ and $\triangle CED \cong \triangle CEB$.
     - Therefore, $BC = CD = AB$.
   - Thus, $AB = BC = CD = DA$, proving that $ABCD$ is equilateral.

FIGURE 15.4 Continued chat with ChatGPT. We have highlighted points in the text with the black boxes.

If the students do not truly understand the earlier output, it could be characterised as justificational mediation, which then turns into a techno-authoritarian proof scheme. ChatGPT-4 relies on the 'side-angle-side (SAS) congruence theorem' and in the first highlighted box in the answer where ChatGPT-4 emphasises triangle ABE and ADE. In the second highlighted box, ChatGPT-4 emphasises triangles BEC and DEC, and CED and CEB, respectively. If the students by themselves notice the differences between the notions in these two boxes, and whether this has a meaning in the argumentation or not, we can value it as a sign of a turnover to an epistemic mediation. In such case, we argue that the students, to some extent, use both their communication competency and their reasoning competency. Here the students might also compare with the text in Euclid's Proposition 6, Book IV, or at least with the illustration. If the students then try to question these differences and relate them to the criteria of ASA and SAS congruence, respectively, they might activate their thinking competency to support their interpretation of a chain of arguments and thus potentially formulate their own arguments and chains hereof, thereby developing their reasoning competency. A teacher could ask the students to look at these two boxes, and ask them: 'What are the similarities and differences between the notions in these two boxes? Which angles and sides are into play here, and how do these correspond with the ASA congruence? How does this correspond with the argument based of a right angle in a semicircle in Euclid's Proposition 6, Book IV?' A teacher might also let the students draw their own construction of triangles by hand or in GeoGebra, or let them use other sources to get information of congruence in order to compare with and explain the ChatGPT-4 answer.

8 Some final remarks

In this chapter, we have tried to unfold why it is important that the students in assessment situations are put into mathematical situations, where they must critically question GAI – here, ChatGPT – output when it seems to be of a justificational character. If the teachers are not aware of this, they risk supporting ChatGPT output function as a justificational mediation, which might turn into a techno-authoritarian proof scheme. Here, the risk is that assessment situations turn into the kind that Jankvist et al. (2021) characterised as number 1 (see Section 4). We find that ChatGPT will allow the implementation of new types of assessment situations, number 2, in which case it is very important that both teachers and students are aware of the convincing linguistic character of ChatGPT's answers. In other words, it is important to support the transition of justificational mediation into epistemic mediation. It appears to be a good idea to let students use ChatGPT as a part of the assessment setup, the kind of assessment situation that Jankvist et al. (2021) characterised as number 3.

We have also tried to focus on the fact that primary historical sources can play an active role in assessment situations where ChatGPT is in play. This is because the teacher and the students can relate the ChatGPT output to the way it is explained in the primary historical source. Hence, the source can support them in questioning ChatGPT, so that they obtain further explanation or, we might say, support them in going into a dialogue with ChatGPT. The historical source can support students having to use both their communication competency and their thinking competency while they follow and produce mathematical arguments and proofs – using their reasoning competency – since the source presents another way of explaining the same piece of mathematics, in our case a mathematical proof. Probably, this will not happen by merely giving the students the source and a task related to this; for example, 'try to explain how you can build a chain of arguments of how Euclid's Proposition 6, Book IV, relies on the five postulates in Book I'. It has to be orchestrated in one way or another. The difficult part seems to be that the teachers do not know what types of justificational situations may occur, because ChatGPT answers may differ according to the questions the students formulate. Hence, it is important to put the students in more open assessment situations, so to speak, where it is obvious to them that they must question ChatGPT answers and must argue why they find the answers reliable or not. The assessment situations must encourage the students' critical thinking and reflective use of ChatGPT. Maybe prefabricated chats with ChatGPT – for example, as the excerpts of such a one that we have presented in this chapter – can be a kind of assessment as well as a way to support the students in developing a reflective and critical use of ChatGPT. Here the teachers could ask the students to reflect on what they think of the prompting, and if they have other suggestions to prompt, which could give them a deeper or new mathematical understanding related to the content in the primary historical source. This could be a way to assess and support the students in developing both their thinking and their communication competencies. The teacher could also ask the students to reflect on the way ChatGPT builds up arguments and maybe relate to the way arguments are built up in the primary historical source. For example, by asking about the differences and the similarities and which chains of arguments the students find most convincing and why. Here, the assessment and the supporting of students' possibilities of developing their reasoning competency are in focus.

Notes

1 We do occasionally refer to 'chat', 'interaction', 'dialogue', and other communication with ChatGPT despite that this is not communication in a strict sense. We do this because these AI tools blur the lines between tool use and communication, and we prefer to maintain this ambiguity in our text.

2 OpenAI. (2024, August). *ChatGPT* [Large language model], Model 4o. https://chatgpt.com/share/55f7893e-13fb-49b3-81aa-c694966ea216

Reference list

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning*, *7*(3), 245–274. https://doi.org/10.1023/A:1022103903080

Eibe, T. (1897a). *Euklids Elementer I–II [A Danish translation of Euclid's elements, book I og II]*. Gyldendal.

Eibe, T. (1897b). *Euklids Elementer III–IV [A Danish translation of Euclid's elements, bog III og IV]*. Gyldendal.
Geraniou, E., Manolis, M., Misfeldt, M., & Jankvist, U. T. (accepted). *From CAS to Generative AI tools in Mathematics Education: Lessons learned and lessons to be learnt*. Presented at TWG-15, CERME-14.
Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange*, *21*(1), 6–13. https://doi.org/10.1007/BF01809605
Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 805–842). Information Age Publishing.
Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal*, *16*(2), 207–223. https://doi.org/10.1080/09585170500136093
Jankvist, U. T., Dreyøe, J., Geraniou, E., Weigand, H.-G., & Misfeldt, M. (2021). CAS from an assessment point of view: Challenges and potentials. In A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalová, & H.-G. Weigand (Eds.), *Mathematics education in the digital age: Learning practice and theory* (1st ed., pp. 99–120). Routledge. https://doi.org/10.4324/9781003137580
Jankvist, U. T., & Misfeldt, M. (2019). CAS assisted proofs in upper secondary school mathematics textbooks. *REDIMAT – Journal of Research in Mathematics Education*, *8*(3), 232–266. https://doi.org/10.17583/redimat.2019.3315
Jankvist, U. T., Misfeldt, M., & Aguilar, M. S. (2019). Tschirnhaus' transformation: Mathematical proof, history and CAS. In I. E. Barbin, U. T. Jankvist, T. H. Kjeldsen, B. Smestad, & C. Tzanakis (Eds.), *Proceedings of the eight european summer university on the history and epistemology in mathematics education ESU 8 | Skriftserien* (pp. 319–330). METU.
Joyce, D. E. (1996a). *Euclid's elements* (Book I). Postulates. http://aleph0.clarku.edu/~djoyce/elements/bookI/bookI.html
Joyce, D. E. (1996b). *Euclid's elements* (Book IV). Proposition 6. https://mathcs.clarku.edu/~djoyce/elements/bookIV/propIV6.html
Lagrange, J.-B. (2005). Using symbolic calculators to study mathematics. The case of tasks and techniques. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators. Turning a computational device into a mathematical instrument* (pp. 113–135). Springer. https://doi.org/10.1007/0-387-23435-7_6
Misfeldt, M., & Jankvist, U. T. (2018). Instrumental genesis and proof: Understanding the use of computer algebra systems in proofs in textbook. In L. Ball, P. Drijvers, S. Ladel, H.-S. Siller, M. Tabach, & C. Vale (Eds.), *Uses of technology in K-12 mathematics education: Tools, topics and trends. ICME-13 monographs* (pp. 375–385). Springer. https://doi.org/10.1007/978-3-319-76575-4_22
Nardi, B. A. (1996). Studying context: A comparison of activity theory, situated action models, and distributed cognition. In B. A. Nardi (Ed.), *Context and consciousness: Activity theory and human – computer interaction* (pp. 69–102). The MIT Press.
Niss, M. A., & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning ideas and inspiration for the development of mathematics teaching and learning in Denmark. IMFUFA tekst no. 485*. Roskilde University. (Published in Danish in 2002). http://milne.ruc.dk/imfufatekster/pdf/485web_b.pdf
Niss, M. A., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics*, *102*(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9
Niss, M. A., & Jensen, T. H. (Eds.). (2002). *Kompetencer og matematiklæring – Ideer og inspiration til udvikling af matematikundervisning i Danmark* [Competences and learning of mathematics – ideas and inspiration for development of mathematics teaching in Denmark]. Undervisningsministeriet. Uddannelsesstyrelsens temahæfteserie nr. 18.

Rabardel, P., & Bourmaud, G. (2003). From computer to instrument system: A developmental perspective. *Interacting with Computers*, *15*(5), 665–691. https://doi.org/10.1016/S0953-5438(03)00058-4
Tamborg, A. L. (2021). Improving mathematics teaching via digital platforms? Implementation processes seen through the lens of instrumental genesis. *ZDM Mathematics Education, 53*, 1059–1071. https://doi.org/10.1007/s11858-021-01282-x
Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x
Thomsen, M. (2022). *Matematikhistoriske originalkilder, ræsonnementskompetence og GeoGebra på mellemtrinnet* [PhD thesis, Historical original sources, reasoning competency and GeoGebra in lower secondary school, AU Library Scholarly Publishing Services]. https://doi.org/10.7146/aul.448
Thomsen, M., & Jankvist, U. T. (2020). Reasoning with digital technologies – counteracting students' techno-authoritarian proof Schemes. In A. Donevska-Todorova, E. Faggiano, J. Trgalova, Z. Lavicza, R. Weinhandl, A. Clark-Wilson, & H.-G. Weigand (Eds.), *Proceedings of the tenth ERME topic conference (ETC 10) on mathematics education in the digital age (MEDA), 16–18 September 2020* (pp. 483–490). Johannes Kepler University.
Thomsen, M., & Jankvist, U. T. (2021). Mathematical thinking in the interplay between historical original sources and GeoGebra. In U. T. Jankvist, A. Clark-Wilson, H.-G. Weigand, R. Elicer, & M. Thomsen (Eds.), *Proceedings of the 15th international conference on technology in mathematics teaching: Making and strengthening "connections and connectivity" for teaching mathematics with technology* (pp. 189–196). https://doi.org/10.7146/aul.452
Trouche, L. (2005). Instrumental genesis, individual and social aspects. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The didactical challenge of symbolic calculators turning a computational device into a mathematical instrument* (pp. 197–230). Springer. https://doi.org/10.1007/0-387-23435-7_9

INDEX