# INVALSI DATA: ASSESSMENTS ON TEACHING AND METHODOLOGIES

IV Seminar "INVALSI data: a research and educational teaching tool"

edited by
Patrizia Falzetti

FrancoAngeli
OPEN ACCESS

INVALSI

INVALSI PER LA RICERCA
STUDI E RICERCHE

## INVALSI PER LA RICERCA

La collana Open Access INVALSI PER LA RICERCA si pone come obiettivo la diffusione degli esiti delle attività di ricerca promosse dall'Istituto, favorendo lo scambio di esperienze e conoscenze con il mondo accademico e scolastico.
La collana è articolata in tre sezioni: "Studi e ricerche", i cui contributi sono sottoposti a revisione in doppio cieco, "Percorsi e strumenti", di taglio più divulgativo o di approfondimento, sottoposta a singolo referaggio, e "Rapporti di ricerca e sperimentazioni", le cui pubblicazioni riguardano le attività di ricerca e sperimentazione dell'Istituto e non sono sottoposte a revisione.



ISBN 9788835131557




ISBN 9788835131557

# INVALSI DATA: ASSESSMENTS ON TEACHING AND METHODOLOGIES

IV Seminar "INVALSI data: a research and educational teaching tool"

edited by
Patrizia Falzetti


FrancoAngeli
OPEN ACCESS

ISBN 9788835131557

Le opinioni espresse nei lavori sono riconducibili esclusivamente agli autori e non impegnano in alcun modo l'Istituto. Nel citare i contributi contenuti nel volume non è, pertanto, corretto attribuirne le argomentazioni all'INVALSI o ai suoi vertici.

*Grafica di copertina*: Alessandro Petrini



ISBN 9788835131557

# *Index*





ISBN 9788835131557

ISBN 9788835131557

# *Introduction*

by Patrizia Falzetti

Didactics is that part of the educational activity and theory which concerns teaching methods, in detail it's the rational organization of methods and actions aimed at obtaining an effective educational project.

The school system has always aimed to achieve quality teaching, which is able, on the one hand, to give adequate responses to the expectations of all the stakeholders and, on the other, to introduce tools, actions, and checks through which the training offer can be constantly improved. This process is undoubtedly linked to scientific research. Researchers and Academics start from the data available to them or collect new ones, to discover and/or interpret facts and to find answers and new cues of reflection. A favorable environment for this work was the Seminar "INVALSI data: a research and educational teaching tool", in its fourth edition in November 2019. The volume consists of six chapters, which are arise within the aforementioned Seminar context and, while dealing with heterogeneous topics, offer important examples of research both on teaching and on the methodologies applied to it. Four of the chapters of this volume, from various points of view, have as object of analysis the questions of the INVALSI tests. In the first chapter, the characteristics of some reflection questions on the language are studied, the elements that influence their difficulty (their placement in the skill levels), and how these questions are solved by students. Chapter two presents the results of a qualitative analysis carried out using TIMSS data: analyzing the open-ended questions.

The authors of chapter four, instead, study differential item functioning (DIF) a bias of a test item, which occurs whenever the probability of response to that item differs between groups of examiners with the same ability level (e.g., groups according to gender, geographical area, etc.).

In chapter six, the authors explain the new procedures that, from the academic year 2018/19, are used for the correction of open-ended questions



ISBN 9788835131557

of the INVALSI tests of Italian and Mathematics, administered in Computer Based mode.

The remaining chapters, three and five, explore two very important topics in the school world: the evaluation and the inclusion.

Evaluation is a pedagogically important and didactically essential topic in the teaching-learning process. The authors of chapter 3 explain the criticism of the Large-Scale Assessment (LSA). They show how attitudes have changed in recent years and how there is now a growing interest in extending the impact of LSA as well as on the evaluation of school systems also in other fields. Chapter five, on the other hand, emphasizes how, in the face of the migratory phenomena that have been affecting Italy in recent years, comparative studies focused on identifying patterns of change in the tolerant attitudes of young people are of great importance.

Tolerance, generally defined as positive feelings toward diversity as well as an understanding and endorsement of equality between different groups (Cote and Erikson, 2009), is considered an important democratic attitude and an essential prerequisite for a peaceful coexistence in the increasingly diverse contemporary societies (Freitag and Rapp, 2015). As the authors write «the monitoring and promotion of tolerance in schools are an essential part of policies focused on inclusive citizenship education and intercultural dialogue».

As a Statistical Service, which for years has taken care of the collection and dissemination of data, we hope that in this, as in the other volumes of the series, the reader will find confirmation of the importance that data play, both in scientific research and in practice in classroom.



ISBN 9788835131557

# *1. Language awareness in the INVALSI tests. Competence levels and students' linguistic reflections*

by Zuzana Toth

From 2018 onwards, the results of the INVALSI tests have been reported in terms of competence levels. The descriptors of the competence levels are identified on the basis of a qualitative analysis of the corresponding test questions. The present study focuses on the descriptors of language awareness and investigates to what extent these descriptors are reflected in students' reasoning on a sample of language awareness questions, administered in the third-year class of lower secondary schools in 2018.

Students' linguistic reasonings were elicited by means of focus-group interviews and analysed by means of qualitative content analysis. The results show substantial differences in students' reasoning on questions representing different competence levels. The main characteristics of students' reasonings are consistent with the descriptors of competence levels developed by INVALSI.

*A partire dal 2018, i risultati delle prove INVALSI sono restituiti in termini di livelli di competenza. I descrittori dei livelli sono identificati in base a un'analisi qualitativa dei quesiti corrispondenti. Il presente contributo prende in esame i descrittori di riflessione sulla lingua ed investiga in quale misura tali descrittori sono riscontrabili nei ragionamenti elaborati dagli studenti su un campione di quesiti di riflessione sulla lingua, somministrati nelle classi III della scuola secondaria di I grado nel 2018.*

*Le riflessioni linguistiche degli studenti sono state elicitate tramite interviste focus-group e analizzate con il metodo di analisi qualitativa del contenuto. I risultati mettono in evidenza la presenza di differenze sostanziali nelle riflessioni degli studenti su quesiti di diverso livello di difficoltà. Tali differenze sono congruenti con la descrizione dei livelli di competenza sviluppata dall'INVALSI.*



ISBN 9788835131557

## 1. Introduction

According to the National Guidelines, developed by the Italian Ministry of Education, language education is an important part of the curriculum of lower secondary schools. Language awareness (henceforth LA) is therefore assessed within the tests of Italian administered by INVALSI, the research institute responsible for the external assessment of learning outcomes in the Italian school system.

This paper is part of a larger study and presents its theoretical background, the methodology of data collection and analysis, and the first, preliminary results, based on the analysis of approximately 10% of the data. Given the paucity of the data, it does not aim to draw any generalisable conclusions.

The study focuses on the LA questions in the tests of Italian administered in the third-year class of lower-secondary schools, corresponding to the 8th year of schooling. From 2018 onwards, the tests have been computer-based, and the results formulated in terms of competence levels. The five competence levels, identified by means of statistical analyses of students' answers (for a detailed discussion see Desimoni, 2018), correspond to five task clusters on a competence scale, where each value represents both the task difficulty and the student ability. Therefore, the description of each competence level is based on a qualitative analysis of the corresponding task cluster, defining what kind of abilities are stimulated by the tasks and "what students *typically* know and can do at given levels of proficiency" (OECD, 2017, p. 276).

In the case of language awareness, the description of competence levels is guided by theoretical concepts such as explicit knowledge about language, implicit linguistic competence, prototypicality, and the distinction between form meaning and function (see Bialystok, 2001; Cenoz, Gorter and May 2017; Lo Duca, 2004, 2018). These concepts are discussed in more detail in the next section, before turning our attention to how they are reflected in students' reasonings.

## 2. Descriptors of competence levels of language awareness

The description of competence levels is mainly based on the identification of elements that influence the the difficulty of the questions and consequently their position on the competence scale. An evaluation of the test results assessed against research on language awareness (Toth, 2019) suggested that the difficulty of LA tasks was influenced by the following factors: 1) the degree of explicitness of analysis required by the question; 2) the com-



ISBN 9788835131557

plexity of form-function relationships in the linguistic element the questions focused on; 3) the prototypicality of the linguistic element. These factors cannot be strictly isolated from each other: the difficulty of the questions and the linguistic analyses stimulated by them result from their interaction. This interaction can be illustrated by contrasting the two questions reported below, which represent the two extremes of our competence scale, the lowest level and the highest.

Question 1.

**Per ognuna delle seguenti situazioni comunicative indica quale tra le due frasi proposte è adatta al contesto.**

a) Il tuo stendipanni si è rotto. Entri in negozio e chiedi:
   1. Buon giorno, vorrei uno stendipanni
   2. Buon giorno, vorrei lo stendipanni
b) La mattina hai ordinato al tuo fornaio di tenerti da parte la tua pizza preferita. A pranzo vai al forno e dici:
   1. Salve, sono venuto a ritirare una pizza
   2. Salve, sono venuto a ritirare la pizza
c) Nella vetrina di un negozio è esposto un solo vestito rosso. Entri e chiedi:
   1. Vorrei provare un vestito rosso che è in vetrina
   2. Vorrei provare il vestito rosso che è in vetrina
d) Vuoi trascorrere una serata con i tuoi amici. Telefoni in pizzeria e dici:
   1. Buona sera, vorrei prenotare un tavolo per quattro persone
   2. Buona sera, vorrei prenotare il tavolo per quattro persone

Question 2.

**Per ognuna delle seguenti frasi indica se il verbo è alla forma attiva o passiva.**

a) I miei genitori vanno spesso alla fiera del libro. Forma attiva/Forma passiva
b) Mio fratello è convocato spesso per le partite in trasferta. Forma attiva/Forma passiva
c) Questi moduli vanno spediti entro la fine del mese. Forma attiva/Forma passiva
d) Dalle Olimpiadi di italiano vengono esclusi gli alunni con un voto inferiore a sei. Forma attiva/Forma passiva
e) Luigi è salito sul treno all'ultimo momento. Forma attiva/Forma passiva
f) Oggi pomeriggio vengono a trovarmi degli amici messicani. Forma attiva/Forma passiva



ISBN 9788835131557

The main differences between these two questions, in terms of degree of explicitness, prototypicality and relationship between form and function are discussed in the next sections.

### *2.1. Degree of explicitness*

Consistent with the idea that explicit and implicit knowledge can mutually influence each other (Ellis, 2017, p. 118), and that language awareness is "partly conscious and partly intuitive" (Svalberg, 2016), the linguistic tasks in the INVALSI tests aim to induce students to exploit both their implicit and explicit linguistic knowledge. Some questions can be answered intuitively, on the basis of students' linguistic sensitivity, while others require a careful selection of linguistic features to focus on, by taking into account various levels of linguistic analysis.

If we locate the language awareness question on a continuum of explicitness, the two questions cited above situate themselves at the opposite ends of the continuum. Question 1 invites students to imagine different communicative situations and choose between two sentences, which differ only in the presence of definite and indefinite article. It does not require an explicit explanation of students' choices; their answers may be intuitive, based on their linguistic sensitivity.

Question 2, on the other hand, focuses on the distinction between the active and the passive voice in a series of sentences. It requires students to take into account different levels of linguistic analysis, such as morphosyntax and semantics; to work with the concept of transitivity; to distinguish between syntactic roles such as subject and direct object, and semantic roles such as agent and patient; and to acknowledge that the transformation from the passive to the active voice (or vice versa) causes a change in the relationship between semantic and syntactic roles.

The results of LA questions suggest that the explicitness of analysis is directly proportional to the question difficulty. In fact, as anticipated, questions one and two are located at the opposite ends of the competence scale. This pattern is consistent with the results of studies on language awareness in English (Myhill, 2000; Watson and Newman, 2017) and Dutch (Van Rijt *et al.*, 2019a; 2019b), which claim that students are more likely to carry out semantic analyses, based on their intuitions around meaning. They have a harder time focusing on morphosyntax, which requires the ability to observe form-meaning relationships, and often implies explicit knowledge about language.



ISBN 9788835131557

### *2.2. Prototypicality of the linguistic element*

A feature connected to the explicitness of analysis is prototypicality. Prototypical elements, which represent "best examples" of the category they belong to (Rosch, 1978), can often be categorised intuitively, while less prototypical elements require a hierarchical view of their properties and a conscious selection of those that are relevant for the analysis.

For instance, in the sentence *Luca mangia la mela* [Luca eats the apple], there is a prototypical subject, represented by a human referent assuming the semantic role of agent and occupying the position of the topic in the sentence. The direct object can also be considered prototypical: it denotes an inanimate referent, assumes the semantic role of patient and occupies the position of comment. If the sentence is transformed into the passive voice (*La mela è mangiata da Luca*), the direct object of the active sentence will be the subject assuming the role of patient in topic position, while the subject of the active sentence will be a complement of agent in comment position. Despite this complex interaction between syntactic roles, semantic roles and sentence information structure, the voice of the sentence can be identified by focusing on the level of semantics, i.e., by asking ourselves whether the constituent in topic position is the agent, a widespread rule of thumb used to identify the subject of active sentences (see Favilla, 2018).

Question 2 cited above, however, focuses on less prototypical sentences. For instance, in sentence *b* [*My brother is often invited to participate in away games*], a passive sentence, the complement of agent is not made explicit, while the subject in topic position denotes a [+human] referent. Thus, the type of subject and the absence of the complement of agent decrease the prototypicality of the sentence and induce the necessity of a syntactic analysis, which takes into account the concepts of verb valency, transitivity and the relationship between semantic and syntactic roles.

Furthermore, as demonstrated by the sentences in Question 2, the complexity of the italian verb paradigm allows for a variety of means to form the passive voice. In addition to using the auxiliary verb *essere* [to be], the verbs *andare* and *venire* can be used both as auxiliary verbs in passive constructions such as *c* and *d*, and with a lexical meaning ([*to go*] and [*to come*] respectively) in active sentences such as *a* and *f*.

To sum up, the prototypicality of a linguistic element seems to be inversely proportional to the difficulty of a question. The reason for this is that prototypical elements share all the defining properties of a category, and allow for intuitive classification, while less prototypical elements require a



ISBN 9788835131557

deliberate selection of properties relevant for the analysis, associated with various levels of linguistic analysis.

### *2.3. Relationship between forms and functions*

Reflecting on the relationship between form and function is a fundamental metalinguistic activity (Van Rijt and Coppen, 2017). This relationship involves a variable degree of complexity. In some cases, there is a one-to-one relationship between form and function. However, more often, the same linguistic form has more functions and, vice versa, the same function can be fulfilled by several linguistic means. The difficulty implied by complex form-meaning relationships is widely discussed in the literature on second language acquisition (see DeKeyser, 2016), but is often overlooked by research on language awareness in the L1. However, some indications can be deduced from the studies by Lo Duca and her collaborators on the classification of words into parts of speech (e.g., Lo Duca and Polato, 2010; Lo Duca *et al.*, 2011). These studies show that students tend to classify words according to semantic criteria, assuming a one-to-one relationship between the meaning of the word and the part of speech it belongs to. For instance, nouns and verbs are defined respectively as words indicating persons and objects are classified as nouns, those indicating processes are classified as verbs. Students struggle with the classification of words that do not fit these criteria, such as nouns denoting processes (e.g., vittoria [victory]), because their classification requires a hierarchical view of properties related to different levels of linguistic analysis, such as morphosyntax and semantics.

If we compare the two questions cited above, it is evident that the distinction between the active and the passive voice involves significantly more complex form-function relationships than the choice between definite and indefinite article. In fact, the use of definite and indefinite article (Question 1) is based on the identifiability of the referent (Grandi, 2010), while the distinction between the active and the passive voice (Question 2) involves complex relationships between syntactic and semantic roles, as well as a variety of form-meaning connections.

To sum up, linguistic elements displaying complex form-function relations are more difficult to analyse due to the following reasons: 1) students tend to focus on the level of semantics because intuitions on meaning are easier to capture than morphosyntactic patterns; 2) the tendency to focus on meaning and semantics makes it more difficult to develop a hierarchical view of properties inherent to different levels of linguistic analysis.



ISBN 9788835131557

## 3. The present study

As exemplified by the comparison of a level 1 with a level 5 question, the three criteria used to characterise competence levels (i.e., degree of explicitness, prototypicality and complexity of form-function relationship) are not strictly separable and interact with each other. Questions requiring less explicit analysis can usually be answered on the basis of semantic or pragmatic intuitions. They focus on prototypical linguistic structures, or linguistic phenomena characterised by linear form-function relationships, such as the use of definite and indefinite article for marking definiteness and indefiniteness. On the other hand, questions requiring more explicit analysis tend to focus on less prototypical forms and structures, and phenomena that involve complex form-function relations, such as diathesis. Thus, a qualitative analysis of LA questions suggests that their location on the competence scale is influenced by the interaction of the three criteria discussed here.

However, the above conclusions are drawn from a comparison of the test results against the studies on language awareness, without any qualitative data about the characteristics of students' linguistic reasonings. Therefore, the present study aims to examine how students solve LA questions when working in small groups and to what extent the above identified criteria are observable in their reflections. The main research question, *What are the main characteristics of students' linguistic reasonings?* can be broken down into two subordinate questions:

1) What are the main characteristics of students' reasonings on a level 1 question compared to a level 5 question?
2) To what extent are these characteristics consistent with the descriptors of the competence levels?

### *3.1. Data collection*

The data analysed here are drawn from a larger study, which also aimed to examine to what extent some modifications along the dimensions of explicitness, prototypicality and form-function complexity affect students' approach to the questions, and to what extents their reasoning varies in relation to school grades. The data collection was therefore carried out in three classes: a third year class of a lower secondary school, as well as a first year and a second year class of an upper secondary school, with a total number of 49 students. The present study focuses on the data collected in the third year class of a lower secondary school, from 18 students. They were interviewed



ISBN 9788835131557

in 5 groups of three or four persons, selected by their teacher, who was asked to put together groups of students with a similar level of language awareness, in order to avoid discussions dominated by the most competent person in the group.

Students were given approximately 15 minutes to think about a series of 13 questions individually, after which they participated in an open-ended, semi-structured group discussion, led by the researcher. Group discussion was preferred to personal interviews because this elicitation technique allows the moderator to remain in the background and take a less active role in the construction of meaning, by encouraging students to discuss the questions with their peers (Van Peer, Hakemulder and Zyngier, 2012, pp. 107-109). The group discussions took place during a regular school day, and were led by the researcher, who was not part of the school staff and had never met the students before. The discussions lasted between 40 and 70 minutes and were carried out in one session per group.

The present study examines the students' reflections on Question 1 and 2, as well as their modified version reported in appendix. The modification of Question 1 aimed to increase its difficulty by removing indications regarding the linguistic context and introducing abstract nouns such as *peace* and *justice.* The modification of Question 2, on the other hand, aimed to decrease its difficulty by directing the students' attention to the possibility of using the verb *andare* both as an auxiliary verb and as a verb with lexical meaning.

The design of the group discussion was informed by explicitation interview techniques (Maurel, 2009; Vermersch, 2014), in the sense that the moderator adopted an open, listening attitude, by giving open prompts such as *Would you please discuss what is the answer to this question? Why did you choose this answer? Why did you exclude the other options?*, etc. More explicit interventions (such as requests to manipulate data or summarise conclusions and formulate hypotheses) were only made in cases where students appeared to be stuck on a problem and not being able to move the conversation forward, or they volunteered their answers without discussing the reasons behind their choices. As suggested by Van Peer *et al.* (2012, p. 109), these interventions aimed to «challenge participants, tease out details, make sure meanings are understood and shared», without pressuring the participants or leading their answers in any particular direction.



ISBN 9788835131557

### *3.2. Data analysis*

The group discussions were audio-recorded and transcribed by the researcher. The transcriptions followed the conventions elaborated within the project Voice (2007), with some modifications. The data were coded deductively (Mayring, 2014, pp. 79-88). The categories were defined on the basis of studies on language awareness and students' linguistic reasoning (e.g., Bialystok 2001; Lo Duca 2004; Toth 2019; Van Rijt *et al.*, 2019a). The formulation of categories was mainly guided by the distinction between explicit knowledge about language and implicit linguistic competence, and the concept of focused attention.

The first stage of coding was carried out by two researchers independently. After coding 25% of the material, the whole category system was revised, by comparing the two versions of coding and discussing differences until an absolute agreement was reached. Subsequently, the whole dataset was coded by the researcher.

The code system used for the analysis is reported in Tab. 1, and contains two code-types: evaluative codes referring to the correctness of students' reasoning and descriptive codes referring to the way students analyse linguistic data.

*Tab. 1 – The code system*

| |
|---|
| Correct solution |
| Incorrect solution |
| Conflicting answers |
| Intuitive manipulation of data |
| Intuitive metalinguistic analysis |
| Rules of thumb |
| Focus on morphosyntactic features |
| Focused manipulation of data |
| Reference to metalinguistic knowledge |
| Focus on meaning and semantics |
| Lack of explicit knowledge |
| Unclear focus |
| Analysis not made explicit |

The first descriptive codes reported in Tab. 1 (i.e. *intuitive manipulation of data* and *intuitive metalinguistic analysis*) contain extracts where students analyse data without making reference to an abstract representation of linguistic structures or explicit knowledge about language. Instead, they seem to follow their implicit linguistic competence and some intuitions induced



ISBN 9788835131557

by the context. This type of reasoning is exemplified in Excerpt 1, where the students explain the choice of the definite article in the sentence *Dopo un lungo periodo di guerre e rivolte, nella regione finalmente regna la pace e la giustizia* [*After a long period of war and rebellion, peace and justice rule the region*]. The students seem to have an implicit awareness of the inherent definiteness of the entities *peace* and *justice*, and the consequent necessity to use the definite article (see Grandi, 2010). Student 21 relies on their implicit competence, while Student 31 carries out an intuitive metalinguistic analysis:

Excerpt 1[1].
**Student 21**: perché cioè non si dice cioè […] regna una pace e una giustizia (.) cioè si dice regna LA pace e la giustizia.
*Because you cannot say […] 'regna una pace e una giustizia' (.) you have to say 'regna LA pace e la giustizia'.*
**Student 31**: è la pace: cioè: la giustizia c'è solo una.
*It is the peace, I mean, there is only one peace.*

In some cases, students apply rules of thumb, i.e., simplistic definitions applied mechanically, without a careful examination of linguistic data. The most frequently used rule of thumb observed in the present study is the following: the transformation of a sentence from the passive to the active voice (or vice versa) requires the inversion of the sentence constituents around the verb. This is observable in Excerpt 2, where the student is trying to transform the sentence *d* of Question 2 into the active voice.

Excerpt 2.
**Student 21**: gli alunni con un voto inferiore al sei sono esclusi dalle Olimpiadi di Italiano.
*The students with a mark lower than six are excluded from the Olympics of Italian language.*

Contrary to intuitive approaches, extracts classified as *focus on morpho-syntactic features*, *focus on meaning and semantics* and *focused manipula-*

[1] In each excerpt, the original numeration of the students is maintained, as it appears in the transcriptions, in order to facilitate the location of the excerpt within the database. Each student is given a code composed of two numbers. The first indicates their number within the group, while the second indicates their group. For instance, student 1 from group 3 is reported as students 13. The excerpts are reported in italian, in the same form as they appear in the transcript (for a detailed description of transcription conventions see Voice, 2007). In addition, a meaning-based translation into English is provided for each excerpt. Features related to intonation, pronunciation, pauses, repetition, self-correction, etc. were not marked in the translations.



ISBN 9788835131557

*tion of data* show that students deliberately select the linguistic features to focus on. In some cases they develop an abstract representation of linguistic data, and are also able to alternate their attention between various levels of linguistic analysis, such as morphosyntax and semantics. This is observable in Excerpt 3, where the student explains why the sentence *b* of Question 2 is passive, and how it can be transformed into the active voice.

Excerpt 3.
**Student 23**: perché: il fratello non compie l'azione?
*Because the brother does not complete the action.*
[…]
**Student 23**: ma. ehm. c'è qualcun altro che compie l'azione che subisce il fratello=
*There is somebody else that completes the action undergone by the brother.*
**Moderator**: =hm e come sarebbe questa frase nella forma attiva?
*How would you transform this sentence into the active voice?*
**Student 23**: allora. mancherebbe il complemento d'agente che diventerebbe soggetto? cioè ad esempio: ehm. l'allenatore? […] ha convocato spesso mio fratello per le partite in trasferta.
*The complement of agent would be missing it would become subject? For instance the trainer? […] often invited by brother to participate in away games.*

Finally, the last three codes contain extracts when students' reflections clearly testified to a lack of explicit knowledge or abstract representation of the data, or extracts with unclear focus, which suggest that students were not able to move the discussion forward. For instance, a student, who claims that the sentence *Luigi è salito sul treno all'utlimo momento [Luigi got on the train at the last moment]* is an active sentence, is asked if it can be transformed into the passive voice. Their answer (Excerpt 4) clearly shows a lack of abstract representation of the sentence structure and of the concept of transitivity, given that he/she transforms the sentence by replacing the intransitive verb with a causative structure.

Excerpt 4.
**Student 21**: allora all'ultimo momento (1) […] il treno ha fatto salire Luigi.
*So at the last moment the train let him board.*

## 4. Discussion

One of the most evident differences between the students' reasonings on the two types of questions concerns their degree of correctness. As shown in



ISBN 9788835131557

Tab. 2, incorrect solutions are almost exclusively related to the level 5 questions (12 out of 13). In addition, 24 out of 25 conflicting solutions, i.e., when students do not agree on the answer, are related to the level 5 questions. These results confirm the high accessibility of level 1 questions to the students.

*Tab. 2 – Occurrence of the codes in the data*

| | *Level 1 questions (article)* | *Level 5 questions (diathesis)* |
|---|---|---|
| Correct solution | 9 | 27 |
| Incorrect solution | 1 | 12 |
| Conflicting solutions | 1 | 24 |
| Intuitive manipulation of data | 1 | 8 |
| Intuitive metalinguistic analysis | 27 | 2 |
| Rules of thumb | 0 | 13 |
| Focus on morphosyntactic features | 0 | 17 |
| Focused manipulation of data | 0 | 11 |
| Reference to metalinguistic knowledge | 0 | 1 |
| Focus on meaning and semantics | 0 | 24 |
| Lack of explicit knowledge | 0 | 8 |
| Unclear focus | 1 | 6 |
| Analysis not made explicit | 0 | 7 |

In order to gain a deeper understanding of how students approach these two question types, Tab. 2 also illustrates how the different types of reasoning are related to level 1 and level 5 questions, which is further discussed in the next two sections.

### *4.1. Reasonings for level 1 questions*

The data reported in Tab. 2 show that students follow an intuitive approach when answering level 1 questions (28 out of 29 excerpts). The modified version of Question 1 (reported in Appendix 1) appears to be slightly more difficult than the original: item d gives origin to some conflicting solutions and one incorrect solution. However, in the majority of cases, the intuitive approaches adopted by the students seem to be accurate enough to provide a correct answer. As exemplified by Excerpts 4 and 5, students perceive the relationship between the use of article and the identifiability of the referent thanks to the linguistic context (Excerpt 4), or due to its uniqueness (Excerpt 5).



ISBN 9788835131557

Excerpt 4.

**Student 22**: perché: nella frase dice la (.) la mattina hai ordinato il tuo fornaio di tenerti da parte la tua pizza preferita (.) a pranzo vai al forno e dici salve sono venuto a ritirare la pizza ma n. non a ritirare Una pizza.

*Because the sentence says in the morning you asked your baker to set aside your favorite pizza for you. During lunchtime you go to the bakery and say I came to pick up the pizza not a pizza.*

**Student 12**: perché ce n'era una in particolare.

*Because there was one in particular.*

Excerpt 5.

**Student 21**: perché cioè non si dice cioè dopo un lungo periodo di guerre e rivolte, nella regione finalmente regna una pace e una giustizia. cioè si dice regna LA pace e la giustizia.

*Because you cannot say after a long period of war and rebellions [indefinite article] peace and [indefinite article] justice rule the region. You have to say [definite article] peace and [definite article] justice rule [over the region].*

**Student 31**: è la pace: cioè: la giustizia c'è solo una.

*It is the peace, I mean, there is only one peace.*

**Student 21**: esatto. non posso dire regnano due paci.

*Exactly, I cannot say two peaces rule [over the region].*

To sum up, consistent with the descriptors of competence level 1, students do not demonstrate abstract reasoning or explicit knowledge about language. They are able to infer the definiteness of the referent and select the appropriate article thanks to their metalinguistic intuitions and implicit competence.

### *4.2. Reasonings for level 5 questions*

Contrary to level 1 questions, level 5 questions induce various types of reasoning, such as intuitive approaches (10), application of rules of thumb (13), focus on morphosyntactic features and manipulation of data (17+11), focus on meaning and semantics (24). These reflections do not always lead to correct answers. As Tab. 2 illustrates, episodes of incorrect or conflicting solutions (12+24) outnumber correct ones (27).

Correct solutions are often associated with a multi-layered understanding of the concept of diathesis, i.e. students take into account both semantic and morphosyntactic features. They develop an abstract representation of the sentence structure, take into account the distinction between semantic and syntactic roles and their complex interaction in passive and active sentences, as exemplified in Excerpt 6, where students comment on sentence *a* from Question 1.



ISBN 9788835131557

Excerpt 6.

**Student 13**: ok. allora. io nella: A ho messo che: il verbo è alla forma attiva?

*So in sentence A I indicated that the verb is in the active voice.*

[…]

**Student 13**: […] perché è il: ehm soggetto che compie questa: che compie: […] l'azione.

*Because it is the subject that undertakes the action.*

[…]

**Student 13**: […] e inoltre perché spesso comunque nelle forme passive c'è un ehm. un complemento d'agente che qui che praticamente ehm che tu puoi. quando si trasformerà alla forma attiva sarà lui il soggetto mentre qua non c'è un complemento d'agente. anche se. sì esistono dei casi che potrebbe essere cioè potrebbe anche non esserci il complemento d'agente […] comunque cioè si capisce che il soggetto è quello che compie l'azione.

*In addition, the passive sentences often contain a complement of agent, which will be the subject if you transform it into the active voice. Even if there are cases when there is no complement of agent […] it is clear that the subject undertakes the action.*

As observed in Excerpt 6, Student 13 alternates their attention between semantic and syntactic features. For instance, he/she makes a distinction between the syntactic categories of subject and complement of agent, and the semantic role of agent. However, this kind of reasoning is fairly exceptional. Several students show a tendency to direct their attention to the level of meaning and semantics, while neglecting morphosyntactic features. For instance, they seem convinced that subject as agent is associated with the active voice, while subject as patient is associated with the passive voice, observable in Excerpts 7 and 8.

Excerpt 7.

**Student 12**: [la frase è passiva] perché anche in questo caso non è il soggetto i moduli che decidono di compiere l'azione.

*The sentence is passive because the forms do not undertake the action.*

Excerpt 8.

**Student 32**: perché: allora il soggetto è gli alunni con un voto inferiore a sei. loro non compiono questa azione ma. ma la subiscono dalle olimpiadi di italiano che è complemento d'agente.

*Because the subject is the students with a grade lower than six. They do not undertake this action but they undergo it. The Italian language Olympics is the complement of agent.*

The strong focus on the level of semantics in these reasonings suggests that students overlook morphosyntactic aspects[2]. In fact, in one of the

[2] In Excerpt 8, poor attention to morphosyntax is also deducible from the misclassification of an indirect complement as a complement of agent.



ISBN 9788835131557

groups, after a short discussion about the possible existence of a relationship between diathesis and verb conjugation initiated by a student, Student 12 explicitly claims that he/she focuses primarily on meaning when working with the concept of diathesis. As shown in Excerpt 9, he/she defines the passive voice as a sentence where the [syntactic] subject undergoes an action and erroneously concludes that verbal morphology is not relevant to determine voice.

Excerpt 9.
**Student 12**: […] perché per me cioè la forma passiva è quando subisci. non c'entra molto […] come viene coniugato il ver <1> bo <1> quindi: […] basta. che per la forma passiva il soggetto subisca.
*For me the passive form is when you undergo [an action]. It is not so relevant how the verb is conjugated. It is sufficient if the subject undergoes an action.*

Several interaction segments show that an exclusive focus on semantics may be misleading, when it is not integrated with an abstract representation of the sentence structure. In fact, when the same group is asked to formulate a passive sentence on their own, Student 12 proposes an active sentence with a non-agentive subject, and proposes a transformation into the active voice by replacing the verb with one requiring an agentive subject, as observable in Excerpt 10:

Excerpt 10.
**Moderator**: […] provate a farmi un altro esempio di una frase passiva
*Try to formulate an example of a passive sentence.*
Student 12: ehm. Marco ha ricevuto un pugno.
*Marco received a punch.*
[…]
**Moderator**: capito. […] e: come lo trasformeresti al? alla forma attiva?
*I understand. And how would you transform it into the active voice?*
[…]
**Student 12**: hanno dato un pugno a Marco.
*They gave Marco a punch.*

The lack of attention to morphosyntactic features, and the consequent lack of abstract representation of the sentence structure, may also lead students to apply rules of thumb. In fact, they often try to transform sentences from the passive to the active voice (or vice versa) by simply inverting the sentence constituents around the verb (as exemplified in in Excerpt 2).

On the other hand, the attention to morphosyntactic features does not guarantee the identification of the correct answer. For instance, some stu-



ISBN 9788835131557

dents conclude that the modified version of Question 2 does not have a correct answer, because all the sentences contain *andare*, an intransitive verb.

To sum up, students' reasonings seem to confirm that the level 5 questions analysed here require a multi-layered understanding of a linguistic phenomenon. Correct answers reflect students' ability to switch their attention back and forth between various levels of linguistic analysis and deliberately select the linguistic features to focus on. On the other hand, incorrect answers are associated with a strong focus on some linguistic features, while overlooking others.

## 5. Conclusion

The data presented here are limited, and do not allow for drawing generalised conclusions. However, they are useful for formulating preliminary conclusions, which can be further investigated by the analysis of the whole dataset.

Consistent with the description of competence levels provided by INVALSI, students reasonings on level 1 and level 5 questions show substantial differences. When working on level 1 questions, all the students follow their implicit competence and linguistic intuitions, without explicitly referring to abstract features such as definiteness. In the great majority of cases, these intuitive reasonings lead them to answer the questions correctly.

Students' reasonings on level 5 questions are more diversified. In addition to following their intuition, they also try to develop an abstract reasoning by focusing on morphosyntactic features and referring to their explicit knowledge about language. However, only in 27 out of 63 cases their reasoning leads to a correct solution. These reasonings demonstrate a multi-layered understanding of the concept of diathesis, as exemplified in Excerpts 3 and 6, where students refer to an abstract representation of the sentence structure and distinguish between syntactic and semantic concepts. Reasonings leading to incorrect or conflicting solutions (36 out of 63) seem to put an excessive emphasis on one type of linguistic feature (often related to the level of semantics), while overlooking others.



ISBN 9788835131557

## Appendix 1

**Modified version of Question 1 (level 1)**

Per ognuna delle seguenti coppie di frasi indica quella corretta dal punto di vista grammaticale.

a)

1) Vale la pena di vedere la mostra sugli impressionisti in via Roma: ci sei andato?
2) Vale la pena di vedere una mostra sugli impressionisti in via Roma: ci sei andato?

b)

1) All'inizio dell'anno accademico, il rettore dell'Università di Padova ha salutato gli studenti.
2) All'inizio dell'anno accademico, un rettore dell'Università di Padova ha salutato gli studenti.

c)

1) Dopo un lungo periodo di guerre e rivolte, nella regione finalmente regna la pace e la giustizia.
2) Dopo un lungo periodo di guerre e rivolte, nella regione finalmente regna una pace e una giustizia.

d)

1) Dalle notizie non si capisce se il ragazzo è colpevole o è vittima della giustizia sommaria.
2) Dalle notizie non si capisce se il ragazzo è colpevole o è vittima di una giustizia sommaria.

**Modified version of Question 2. (level 5)**

In quale delle seguenti frasi il verbo *andare* è alla forma passiva?

A. I miei genitori sono andati alla fiera del libro ogni anno.
B. Il viaggio è andato bene, nonostante il maltempo.
C. Tutta la biblioteca è andata distrutta nell'incendio.
D. Due anni fa il paese è andato incontro a una crisi economica.



ISBN 9788835131557

## References


Bialystok E. (2001), *Bilingualism in Development: Language, Literacy, and Cognition*, Cambridge University Press, Cambridge.

Cenoz J., Gorter D., May S. (*eds.*) (2017), *Language awareness and Multilingualism*, vol. 6 of *Encyclopedia of Language and Education*, Springer, New York, 3rd ed.

DeKeyser R.M. (2016), "Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down", *Studies in Second Language Acquisition*, 38, 2, pp. 353-363.

Desimoni M. (2018), *I livelli per la descrizione degli esiti delle prove INVALSI. Le rilevazioni degli apprendimenti (a.s. 2017-2018)*, retrieved on April 6, 2021, from: https://INVALSI-areaprove.cineca.it/docs/2018/Livelli_INVALSI_g8.pdf.

Ellis N.C. (2017), "Implicit and Explicit Knowledge about Language", in J. Cenoz, D. Gorter, S. May (*eds.*), *Language awareness and Multilingualism*, vol. 6 of *Encyclopedia of Language and Education*, Springer, New York, 3rd ed., pp. 113-124.

Favilla M.E. (2018), "Colui, colei o l'oggetto che compie un'azione" Caricature, semplificazioni e stereotipi nell'apprendimento di una nozione sfuggente, in E. Calarescu, S. Dal Negro (a cura di), *Attorno al soggetto. Percorsi di riflessione tra prassi didattiche, libri di testo e teoria*, Studi AitLA, Bologna.

INVALSI (2018), *Quadro di riferimento delle prove INVALSI di Italiano – Documento*, retrieved on April 6, 2021, from: https://INVALSI-areaprove.cineca.it/docs/file/QdR_ITALIANO.pdf.

Lo Duca M.G. (2004), *Esperimenti grammaticali. Riflessioni e proposte sull'insegnamento della grammatica dell'Italiano*, Carocci, Roma.

Lo Duca M.G. (2018), "Le prove di grammatica dell'INVALSI e la progressione dei contenuti grammaticali: il caso del soggetto", in E. Calarescu, S. Dal Negro (a cura di), *Attorno al soggetto. Percorsi di riflessione tra prassi didattiche, libri di testo e teorie*, AItLA, Milano, pp. 123-138.

Lo Duca M.G., Cristinelli A., Martinelli E. (2011), "Riconoscere le voci verbali: indagine su una categoria complessa", in L. Corrà, W. Paschetto (a cura di), *Grammatica a scuola*, FrancoAngeli, Milano, pp. 153-171.

Lo Duca M.G., Polato S. (2010), "Dalle elementari alle soglie dell'università: indagine sul riconoscimento della categoria lessicale del nome", in G. Fiorentino (a cura di), *Perché la grammatica? La didattica dell'italiano tra scuola e università*, Carocci, Roma, pp. 78-92.

Maurel M. (2009), "The Explicitation Interview: Examples and Applications", *Journal of Consciousness Studies*, 16, pp. 58-89.

Mayring P. (2014), *Qualitative content analysis: theoretical foundation, basic procedures and software solution*, SSOAR, Klagenfurt.

Myhill D. (2000), "Misconceptions and Difficulties in the Acquisition of Metalinguistic Knowledge", *Language and Education*, 14, 3, pp.151-163.

OECD (2017), *PISA Technical report 2015*, retrieved on April 6, 2021, from: https://www.oecd.org/pisa/data/2015-technical-report/.




ISBN 9788835131557

Svalberg A.M-L. (2016), "Language Awareness", in G. Hall (*ed.*), *The Routledge Handbook of English Language Teaching*, Routledge, New York, pp. 399-412.
Toth Z. (2019), *La descrizione dei livelli di competenza in base alla prova INVALSI di Italiano*, Working Papers INVALSI 39.
Van Peer W., Hakemulder F., Zyngier S. (2012), *Scientific Methods for the Humanities*, John Benjamins, Amsterdam & Philadelphia.
Van Rijt J.H.M., De Swart P., Wijnands A., Coppen P.A.J.M. (2019a), "When students tackle grammatical problems: Exploring linguistic reasoning with linguistic metaconcepts in L1 grammar education", *Linguistics and Education*, 52, pp. 78-88.
Van Rijt J.H.M., Wijnands A., Coppen P.A.J.M. (2019b), "How secondary school students may benefit from linguistic metaconcepts to reason about L1 grammatical problems", *Language and Education*, 34, 3, pp. 231-248.
Vermersch P. (2014), *The explicitation interview*, retrieved on April 6, 2021, from: https://www.academia.edu/36572134/The_explicitation_interview.
VOICE Project 2007, *VOICE Transcription Conventions* [2.1], full text retrievable from: https://www.univie.ac.at/voice/page/transcription_general_information.
Watson A.M., Newman R.M.C. (2017), "Talking grammatically: L1 adolescent metalinguistic reflection on writing", *Language Awareness*, 26, 4, pp. 381-398.



ISBN 9788835131557

# *2. TIMSS 2015: focus on mathematics errors in open-ended questions*

by Francesco Annunziata, Laura Palmerio

The TIMSS study (Trend in International Mathematics and Science Study) promoted by IEA aims at measuring student educational achievement in Mathematics and Science at 4th and 8th grade. The four-year study frequency, with Italy's participation since the first cycle, enables the study of the trends highlighting, the development of the students' achievement from 4th to 8th grade. In the present work, we intend to focus on the Italian 8th graders results concerning Math questions of TIMSS 2015. While the quantitative analysis of the answers gives us a general context of the results of italian students' achievement, the use of a qualitative approach aims at deepening the analysis of answers given by students detecting additional features not reported. Our choice to analyze Mathematics questions is due to the possibility to compare them to national assessments' questions, since Mathematics in 8th grade is one of the assessed subjects. Based on TIMSS 2015 report definitions, we chose to focus our analysis on the cognitive processes of *Applying* and *Reasoning*. In more than half of the questions belonging to those two processes, the correct answer percentage is below 50%. Among those questions we opted for the analysis of open-ended questions in order to go beyond the mere judgment of correct or incorrect answer and comprehend the resolution strategies underlying the students' answers and the possible reasons for the error. The choice of the questions to analyze was based on coding criteria for the open-ended items, that is to say the classification process into predefined categories which led to a certain scores assignment. The choice was then oriented to items for which a significative number of codes for incorrect answers were defined, in order to have a more detailed starting framework of the reasons underneath the error and the more common misconceptions. In the analysis, the incorrect answers have been divided into additional conceptual categories, inside which we operated another classifica-



ISBN 9788835131557

tion based on the type of error. The results are presented taking into account the italian geographical macro-areas of the sampled schools and the gender of the student, to evaluate if there are any significant differences between those categories and to provide suggestions related to subject teaching.

*L'indagine TIMSS (Trends in International Mathematics and Science Study) promossa dalla IEA ha come obiettivo la rilevazione degli apprendimenti degli studenti in Matematica e Scienze al quarto e all'ottavo grado di scolarità. La frequenza quadriennale dell'indagine, con la partecipazione dell'Italia fin dal primo ciclo, permette di studiarne i trend, evidenziando l'evoluzione nel tempo dei risultati della stessa coorte di studenti dal quarto all'ottavo grado. In questo studio è stata effettuata un'analisi di tipo qualitativo dei risultati della prova cognitiva di Matematica della rilevazione TIMSS 2015 conseguiti dagli studenti italiani dell'ottavo grado. Se da un lato le analisi quantitative già disponibili nel rapporto nazionale TIMSS 2015 permettono di avere un quadro generale sugli apprendimenti degli studenti italiani, dall'altro il lavoro di analisi qualitativa consente di approfondire le risposte fornite dagli studenti, evidenziandone caratteristiche differenti da quelle fino ad ora trattate. Si è scelto di esaminare la Matematica all'ottavo grado anche per avere un possibile confronto con le prove standardizzate nazionali, essendo una delle materie oggetto di valutazione, insieme all'Italiano e all'Inglese. Sulla base delle definizioni del rapporto nazionale TIMSS 2015, abbiamo deciso di circoscrivere la nostra analisi alle domande inerenti ai processi cognitivi di Applicazione e Ragionamento. Per più della metà delle domande totali che rientrano in questi due processi cognitivi, si evidenzia che la percentuale di risposte corrette non raggiunge il 50%. Tra queste si è scelto di analizzare le domande a risposta aperta – che rappresentano il 40% delle domande totali – in quanto permettono di comprendere maggiormente le strategie di risoluzione sottese alle risposte degli studenti e definire con più accuratezza i possibili motivi di errore. La scelta di quali domande aperte prendere in considerazione è stata definita sulla base dei criteri di codifica delle risposte degli studenti alle domande aperte, ovvero il processo di classificazione delle risposte aperte in categorie prestabilite, cui consegue l'assegnazione di un codice utilizzato per attribuire il punteggio. La scelta è stata quindi orientata verso quegli item per i quali è previsto, in fase di codifica, un numero significativo di codici da assegnare alle risposte sbagliate, così da avere un quadro di partenza degli errori più comuni e delle misconceptions. In fase di analisi, le risposte sbagliate sono state suddivise in ulteriori categorie concettuali, all'interno delle quali è stata effettuata una classificazione aggiuntiva in base alla tipologia di errore. I risultati ot-*



ISBN 9788835131557

*tenuti sono stati valutati in relazione alla macro-area geografica nella quale ricadono le scuole campionate e in relazione al genere dello studente, per valutare possibili differenze significative all'interno di queste categorie e cercare di fornire possibili spunti per la didattica della materia.*

## 1. Introduction

The concept of error plays a fundamental role in reflections on teaching and learning mathematics since the observation of errors made by students suggests possible strategies for improvement and new ideas in teaching.

Popper (1972, as cited in Zan, 2007, p. 22) wrote that «avoiding mistakes is a petty ideal: if we do not dare to face problems that are so difficult that error is almost inevitable, then there will be no development of knowledge. In fact, it is from our boldest theories, including the erroneous ones, that we learn the most. No one can avoid making mistakes, the greatest thing is to learn from them». In fact, it is the recognition of the type of error from which a teacher can draw inspiration for targeted teaching.

Russell and Masters (2009), in a paper presented at the annual meeting of the American Education Research Association, noted that when analysing errors, teachers might overlook students' conceptual understanding in favour of a procedural correction. Ketterlin-Geller and Yovanoff (2009, p. 6) also noted that teachers might find it difficult to distinguish between a "lapsus" and a "bug" error. By definition, the analysis of errors made by students has as its main objective the desire to bring out the motivation of the type of reasoning error. The analysis promoted by Ketterlin-Geller and Yovanoff (2009) was concerned with the errors (or "bugs") that students make based on their lack of understanding of the stimulus or procedures to be implemented. The element that emerges from these authors is that mathematical errors occur when the student believes that what has been done is correct, confirming the error by reporting an incorrect reasoning.

To explore the types of errors made by students in mathematics, we used the results of TIMSS 2015.

TIMSS, promoted by the IEA (International Association for the Evaluation of Educational Achievement), aims to assess students' learning in mathematics and science in their fourth and eighth years of schooling.

The first TIMSS survey dates back to 1995, with the participation of Italy since the first cycle. Conducted every four year, TIMSS allows studying trends, examining also changes over time within a cohort of students, given that the cohort of 4$^{th}$ graders in one cycle is assessed again as 8$^{th}$ graders in



ISBN 9788835131557

the next cycle. The sample of students involved in the survey, representative at both the national and macro-geographical level, was extracted following a two-stage stratified sampling procedure: in the first stage, schools were selected with probabilities proportional to their size. In the second stage, one or two intact classes (of grade 4 and/or 8) were randomly selected within the schools sampled in the first stage.

The TIMSS 2015 italian sample consisted of 161 schools and a total of 4,481 students – 2,224 girls and 2,257 boys (INVALSI, 2016). As an international survey, to allow for comparisons between the participating countries and with the previous cycles of the survey, the TIMSS 2015 survey was conducted towards the end of the school year, which coincided with the months of March and April 2015 for countries in the northern hemisphere; for countries in the southern hemisphere, in contrast, the administrations occurred in the period between October and November 2014.

The main objective of the TIMSS is to provide participating countries with a tool to monitor and evaluate the teaching and learning of mathematics and science at different levels of schooling and over time. Specifically, this tool makes it possible to obtain internationally comparable data for the two levels of schooling and to monitor trends in the learning of mathematics and science within each school level under investigation.

In the TIMSS framework, each subject investigated is organised around two dimensions: content and cognitive. The content dimension defines the subject to be assessed within mathematics or science, while the cognitive dimension defines the thinking processes to be assessed.

Concerning 8$^{th}$ grade mathematics, there are four content domains: *Number, Algebra, Geometry, and Data and chance*. Each content domain is composed of different topic areas, as reported in Table 1.

*Tab. 1 – TIMSS grade 8 mathematics content domains and percentage of assessment for each domain*

| *Content domains* | *Topic areas* | *Weighting of TIMSS assessment (%)* |
|---|---|---|
| Number | Integers<br>Fractions and decimals<br>Ratios, proportions and percentages | 30 |
| Algebra | Expressions, operations and equations<br>Relationships and functions | 30 |
| Geometry | Geometric shapes and measurements | 20 |
| Data and probability | Characteristics of data sets<br>Data interpretation<br>Probability | 20 |



ISBN 9788835131557

Three cognitive domains are assessed across and in conjunction with the content domains outlined above: *Knowing*, *Applying* and *Reasoning*. The cognitive domain *Knowing* (35% of the total number of questions), addresses the facts, concepts and procedures that students must know to solve the questions; the domain *Applying* (40% of the questions) refers to the students' ability to apply ideas and conceptual knowledge to solve problems or answer questions; finally, the domain *Reasoning* (25% of the questions) goes beyond solving routine problems to include unfamiliar situations, complex contexts and problems that require a multi-step solutions.

Across the eighth-grade mathematics assessment, each content domain receives approximately equal weight in terms of the numbers of items allocated to assess the topic.

The survey uses a booklet rotation in which each student solves some of the prepared items, respecting this distribution. The TIMSS 2015 test was paper-based and was administered in two sessions of 45 minutes' duration.

TIMSS being a test comprising both multiple-choice and constructed-response items, the aim of this in-depth study was to perform a qualitative analysis of the latter given by Italian eighth grade students.

If the quantitative analyses already included in the TIMSS 2015 national report allow us to obtain a general picture of Italian students' achievement, a qualitative analysis allows us to better explore characteristics of the thinking process that may remain hidden behind the right/wrong coding.

## 2. A brief summary of the mathematics results of Italian students on the TIMSS 2015 in grade 8th

The results of the TIMSS 2015 international survey were analyzed according to distinct perspectives. The first step was to compare the results of Italian students with those of students from other participating countries to consider Italy's positioning in the international arena. Then, the students' performances were analyzed with respect to the content domains and cognitive domains as well as in relation to the four international skill levels (benchmarks) that correspond to four different points of the overall mathematical scale. Finally, a comparison was performed between the Italian macro-areas to evaluate the different levels of students' learning across the national territory.

On the TIMSS 2015, Italy is tested at grade 8 in mathematics, with an average score of 494 points, which is slightly but significantly lower than the international average (500). This result is not significantly different from



ISBN 9788835131557

that in 2011, as occurred for the other 12 countries. Italy consolidated the considerable improvement achieved between 2007 and 2011, which was the greatest among the participating countries.

A substantial gap emerges with regard to the differences in scores between the geographical macro-areas into which the Italian territory is divided: the Northeast stands out positively, with an average score of 520 (significantly higher than the national average of 494), while in contrast, the South-Islands have the lowest score and are significantly lower than the average of Italy, with a value of 452. The macro-areas of the Northwest and the Centre report scores do not differ significantly either from the average of Italy as a whole or from the international average, while the South has a significantly lower score than the international average (484).

As far as gender is concerned, boys score, on average, 7 points higher than girls, but examining the data disaggregated at the macro-geographical area level, the advantage of boys over girls is only confirmed in the South, with a significant difference in the average score of 13. However, in the remaining geographical areas, there is no significant difference between the results achieved by boys and girls. The gender difference is mainly reflected in the content domains.

In Italy, boys significantly exceed girls in the content domains *Number* (+ 19 points) and *Data and chance* (+10 points), while girls exceed boys by 7 points in *Algebra*. Within the macro-areas there are no differences between the two genders for the content domains *Geometry* and *Algebra*, while for domain *Data and chance*, boys score significantly higher than girls in the South (+15 points) and for *Number* in all macro-areas except the South-Islands (Northwest +18 points; Northeast +16 points; Centre +26 points; South +25 points).

In the cognitive domains, in Italy, there are no significant differences between boys and girls except for *Applying*, on which boys exceed girls by 6 points. Across the geographical macro-areas there are generally no gender differences in the various cognitive domains except for the South, where boys score significantly better in the *Applying* domain (+11 points).

Finally, the results were analyzed based on the socio-economic and cultural indicator (SES), which is essentially based on the availability of some resources for study at home and parents' level of education and type of occupation. In general, there is a systematic and positive association between the level of this indicator and the average score in mathematics. In fact, internationally, 13% of students rank high on the indicator and score an average of 540 in mathematics. 72% of students place at the intermediate level of SES and get an average mathematics score of 481. The remaining 15% that are at the low level of SES score only 431 points. In Italy, the percentages of



ISBN 9788835131557

students who fall into the three levels of SES are the same as internationally, and they obtain mathematics scores of 540, 497, and 444, respectively.

As for the geographical macro-areas, in the Northeast and the Centre the percentages of students at the high level of the indicator increase to 19% and 16%, respectively, while in the South-Islands the percentages of students at the high level of the indicator fall to 6%; at the low level of the indicator, in contrast, the South and South-Islands have relatively higher percentages of students, 21% and 27%, respectively.

## 3. Methods

Qualitative analysis of incorrect answers was performed starting with the reading of the answers given by the students sampled from the TIMSS 2015 surveys, which were coded by a team of expert coders in mathematics and were classified as "incorrect responses".

Mathematics at the eighth grade (average score of 494) was examined compared to mathematics in the fourth grade, which is another degree of schooling investigated by the survey. The 8$^{th}$ grade student cohort took the TIMSS test in 2011, during their 4$^{th}$ year of primary school, obtaining an average score of 508 points. Since the average score observed has decreased over the years, the 8$^{th}$ grade TIMSS questions were analyzed.

Based on the definitions of the TIMSS 2015 national report, to correctly answer the questions on the survey, a student must not only be familiar with the contents of the surveyed mathematics but must also demonstrate several cognitive skills. In this work, only questions related to the cognitive processes of *Applying* and *Reasoning* were analyzed.

Subsequently, among the TIMSS 2015 items, the analysis focused on open-ended questions – which represent 40% of the total questions – since they allow for a better understanding of the resolution strategies underlying the students' answers and more accurately define the possible reasons for errors.

The choice of which open-ended questions to consider was defined on the basis of the criteria for coding students' answers to the open-ended questions, i.e. the process of classifying open-ended answers into pre-established categories, followed by the assignment of a code used to assign the score.

The choice was therefore oriented towards the items for which a significant number of codes were assigned to wrong answers during the coding process to have a starting point for the most common errors and misconceptions.

Subsequently, during the analysis phase, the reading of the single answers of the students among the single items selected allowed for new conceptual



ISBN 9788835131557

categories to be explored and developed according to the type of error. In this process, the objective was to identify and organize all the answers provided by the students to explore, in detail, the type of error committed by the students for every single item.

Finally, the results obtained were compared in relation to the student's gender, the geographical macro-areas, and the content domains to evaluate possible significant differences within these categories and to attempt to provide possible cues for teaching.

## 4. Results

To detect students' knowledge and skills in Mathematics on the TIMSS 2015 survey, the items used were constructed (as described above) by associating four content domains with three cognitive processes. Among the various items present on the TIMSS 2015, 100 items require an open answer. The preliminary analysis of this work was to define the overall situation of the students' answers to the questions that required an open answer, thus attempting to understand the percentage of correct answers given for every single item by the students who took the test.

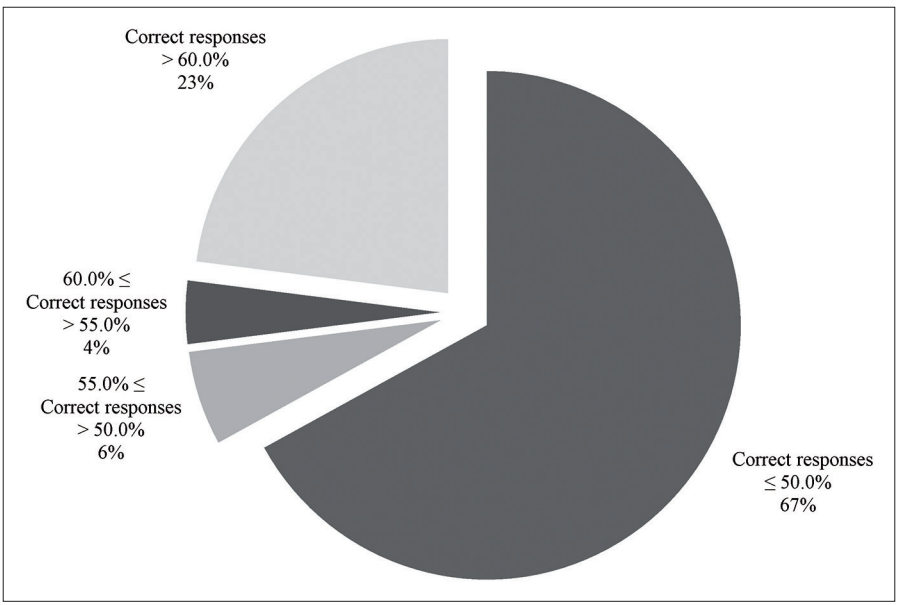


*Fig. 1 – Percentage of open-ended items on the TIMSS 2015 MS*



ISBN 9788835131557

The graph shows that only 23% of the open-ended questions had a "correct answer" coding of more than 60%. More than half of the items, precisely 67%, had a correct answer rate of 50% or less.

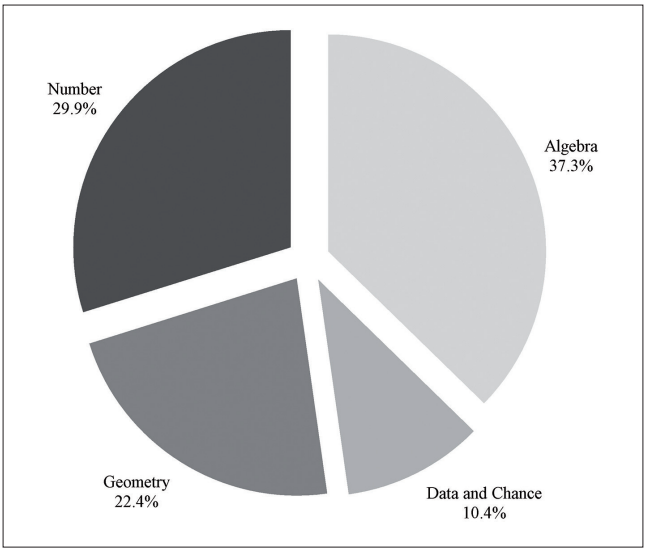


*Fig. 2 – Percentage of content domains with a percentage of correct responses ≤ 50.0%*

Specifically, exploring only the open-ended items that had a percentage of correct answers less than or equal to 50% in detail, it can be seen (Fig. 2) that 37.3% of cases were mainly represented by the *Algebra* content domain, followed by the *Number* (29.9%), *Geometry* (22.4%), and *Data and chance* (10.4%) domains.

After a careful reading of all the items with less than 50% of correct answers, a significant open-ended question for each content domain was selected.

In total, 8 items – 2 for each content domain – were selected and analyzed, considering only the cognitive domains *Applying* and *Reasoning*, with the aim of identifying common errors for each item.

Among the questions that presented these characteristics, one was among the items released, for which it was possible to show and describe the results in an explanatory manner. The other 7 questions, the details of the errors and the stimulus cannot be explained in this paper, but it is possible to present the percentages that emerged from each conceptual category and the questions in a very abstract fashion.



ISBN 9788835131557

*Tab. 2 – Content domain* Algebra

| | | |
|---|---|---|
| | | *Incorrect responses: 457/472* |
| Algebra applying | 2.9% | The students seem to have thought about the question but were not able to answer |
| | 44.6% | The students identified the value of the unknown but were not able to provide an answer |
| | 52.5% | Incorrect responses |
| | | *Incorrect responses: 320/422* |
| Algebra reasoning | 50.9% | The students applied the correct mathematical rule but did not reach the correct solution |
| | 40.9% | The students used another rule that was completely wrong |
| | 8.2% | The students gave vague or incomplete answers |

The questions on *Algebra-Applying* require indicating the greater value of an unknown in an equation and justifying the answer. In general, this type of question presented 3.2% correct answers, 15 out of a total of 472. In 2.9% of the incorrect answers, the students seemed to have thought about the question but were not able to provide an answer; and 52.5% of the incorrect answers incorrectly identified the unknown. In the remaining 44.6% of cases, the students identified the value of the unknown but were not able to provide an answer. As far as the cognitive domain of *Algebra-Reasoning* is concerned, the questions present a numerical sequence with positive and negative numbers, in which, once the rule is established, the missing value must be identified. In 50.9% of cases, the students applied the correct mathematical rule but did not reach the correct solution; among these students, some students, although they obtained the solution, did not consider the presence of the positive/negative sign. In 40.9% of cases, the students used another completely wrong rule, while in the remaining 8.2%, the students gave vague or incomplete answers in an attempt to give a solution but without providing a concrete answer.

In the questions of *Geometry-Applying*, the students are asked to apply the formula of the area of a known solid to an abstract figure. To be able to answer these questions correctly, the student must break the figure down and transform it into the shape of the known solid. The correct answer to this question does not require the support of any type of calculation but a reorganization of the figure, considering the squares present on the sheet on which it is drawn. This item presented 69.7% incorrect answers (303 of 435 total). Specifically, 36% of the students showed knowledge of other geometric rules, inserting formulas or geometric definitions but neglecting the basic principle of the stimulus; 23.1% provided vague answers, and 35% of the an-



ISBN 9788835131557

swers presented only signs on the graph, assuming possible reasoning for the question without providing an answer; and finally, in 5.9% of the answers, the students inserted only the formula for the area of the solid indicated in the stimulus without being able to apply it to answer the question. To represent the cognitive domain of *Geometry-Reasoning*, a question was identified in which, by presenting two solid figures with the same shape and the same dimensions (almost completely overlapping) and highlighting only two specific parts that do not overlap each other, the students are asked to determine the equality between the two areas. This item presented 17.9% correct answers; 16.8% of the 82.1% of students who provided incorrect answers provided a response repeating the question, 6.3% seemed to have thought about the question but were not able to provide an answer and, finally, 76.9% provided vague answers without answering the stimulus.

*Tab. 3 – Content domain* Geometry

| | | |
|---|---|---|
| | | *Incorrect responses: 303/435* |
| Geometry applying | 36.0% | The students wrote generic geometric rules but neglecting the basic principle of the stimulus |
| | 23.1% | The students gave vague or incomplete answers |
| | 35.0% | The students seem to have thought about the question but were not able to answer |
| | 5.9% | The students tried to use the specific geometric rule underlying the question but were not able to concretely apply it in order to provide the correct answer |
| | | *Incorrect responses: 335/408* |
| Geometry reasoning | 16.8% | The students provided a response repeating the question |
| | 6.3% | The students seem to have thought about the question but were not able to provide an answer |
| | 76.9% | The students gave vague or incomplete answers |

In the *Number-Applying* question, the students are asked to determine the major one between two fractions, and to provide the motivation for the answer. In this case, 82.2% of the answers given by the students were classified by the expert coders as incorrect; of these, 91% provided an incorrect answer, 7% provided the correct answer but did not give a correct motivation, and 2% provided a vague or incomplete answer without responding to the stimulus. The *Number-Reasoning* question, in contrast, presents a fractional mathematical problem in which the placement of an initial integer number is followed by a series of actions indicated in a fractional manner, and the students are asked to identify the correct answer and provide the motivation



ISBN 9788835131557

for it. This item presented 17.7% correct answers (104 of 589), of which, in 13.6% of cases, the students provided an incorrect response with or without an explanation and 45.4% provided a vague response or did not understand the question. In 41% of cases, the students marked the correct response but did not provide any explanation (13.2%) or were not able to explain (27.8%).

*Tab. 4 – Content domain* Number

| | | *Incorrect responses: 465/566* |
|---|---|---|
| Number applying | 91.0% | The students provided an incorrect answer |
| | 7.0% | The students provided the correct answer but did not give a correct motivation |
| | 2.0% | The students gave vague or incomplete answers |
| | | *Incorrect responses: 485/589* |
| Number reasoning | 13.6% | The students provided an incorrect response with or without an explanation |
| | 45.4% | The students gave vague or incomplete answers |
| | 13.2% | The students marked the correct response but did not provide any explanation |
| | 27.8% | The students market the correct response but were not able to explain |

*Tab. 5 – Content domain* Data and chance

| | | *Incorrect responses: 522/604* |
|---|---|---|
| Data and chance applying | 81.8% | The students provided a completely wrong answer |
| | 18.2% | The students provided an almost correct answer |

The item chosen for the cognitive domain of *Data and chance-Applying* concerns the representation of data. The students are asked to indicate the correctness or not of a graphical representation of data presented in a table and to provide a justification for their answer. To answer correctly, the students must demonstrate that they understand that the values on the x-axis have been inserted at non-equivalent intervals. Of the students, 86.4%, or 522 of 604, provided an answer coded as an "incorrect response" by the team of experts. In this case, from reading the individual answers obtained by the students, it was possible to categorize the answers into two main groups: the first group consisted of "completely wrong answers," including both wrong answers without motivation (17%) and wrong answers with motivation (64.8%), and the second group consisted of "almost correct answers," in which we instead found correct answers to the first stimulus, but 3.1% of students did not provide a motivation for their answer and 15.1% provided an incorrect motivation. Moreover, the second group of "almost correct an-



ISBN 9788835131557

swers" that had wrong motivations (15.1%) was characterized by motivations concerning the arrangement of the axes rather than the arrangement of the intervals in 49% of cases.

The question on *Data and chance-Reasoning*, reported below, constitutes one of the items released.

The question is based on the representation of data, i.e., starting from the representation of a bar chart, the point of origin of which is different from 0, and the student is asked to provide the motivation for the error in the interpretation of John's results.

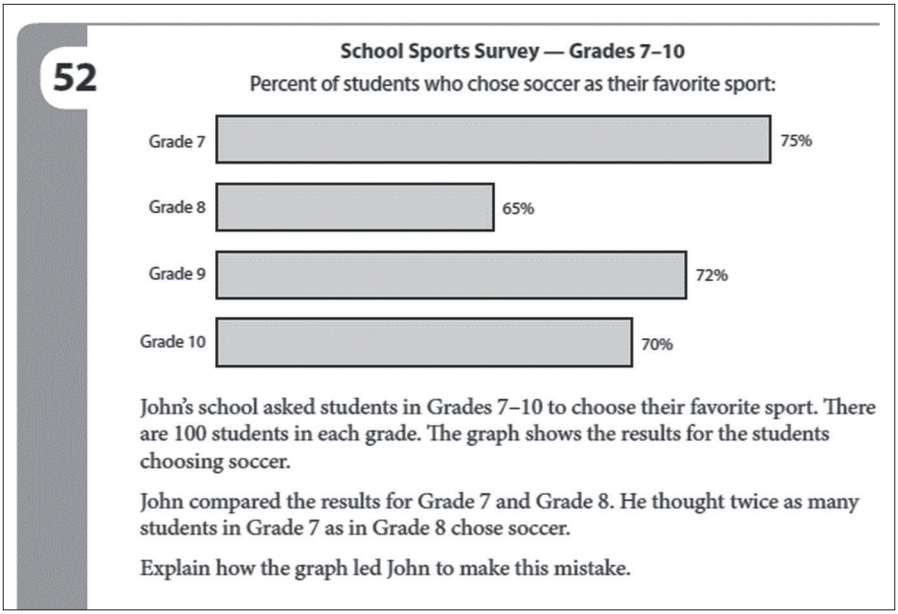


*Fig. 3 – Released item: Data and chance-Reasoning*

The coding guides presented as "correct responses" the answers given by the students in which they indicated the following: "The grade 7 bar is twice as long as the grade 8 bar or equivalent" or "The origin notes t 0" or "The graph is not drawn to scale". In incorrect responses, the answers did not fall within the three options indicated as correct, including crossed out, erased, stray, of illegible marks or off-task responses.

The item had 62% incorrect answers, among which one could create conceptual subcategories. In 32.9% of cases, students provided vague, incorrect answers, such as "John makes this mistake because he sees that grade 7 students have the highest percentage"; they did not refer to the graph or to the



ISBN 9788835131557

concept of "twice"; or they wrote, "He's wrong because he thinks that 65% is twice 75%". The students found the data corresponding to grade 7 (75%) and grade 8 (65%) on the graph and stated that John was wrong without explaining how John made the mistake. In 36.7% of the incorrect answers, the students claimed that the graph was drawn badly: this category included all the answers in which students referred to the data representation without explaining what might be misleading (the graph was not wrong) in the way that the data were represented. Among these students, we found answers such as "because he drew the graph wrong", "he drew the graph wrong in the second and third grade", and "the graph was simply drawn wrong".

Of the incorrect answers, 30.4% were characterized by answers in which the students reasoned on the question, but in 4.6%, they did not answer or did not provide an answer; in 17.5% of the cases, they pointed out that the percentage value was not double but did not respond to the stimulus; and finally, in 8.3% of the cases, the students indicated the percentage difference between the two values indicated but did not report a motivation.

## 5. Other results

The items identified in this work do not allow for inferences to be drawn about national and international relationships because only two items were considered for each single content domain – one for each cognitive domain.

A descriptive analysis of the items under examination reveals slight differences related to gender.

The graph shows that, regarding the cognitive domain of *Applying*, just more than half of the incorrect answers were given by boys (54%) compared to 46% by girls; specifically, out of 100 boys to whom the question was posed, 86 gave incorrect answers in *Applying* compared to 82 girls.

As far as *Reasoning* is concerned, 53% of the wrong answers were given by girls compared to 47% by boys; specifically, of 100 girls to whom the question was attributed, 77 gave incorrect answers compared to 74 boys.

At the macro-area level, on the TIMSS 2015, we can observe that *Reasoning* is the strong point of all the macro-areas, except for the South and the South-Islands, which scored the average, while the cognitive domain *Knowing* is the weak point of the Northern regions. For the cognitive domain *Applying*, there are no significant differences between the macro-areas.



ISBN 9788835131557

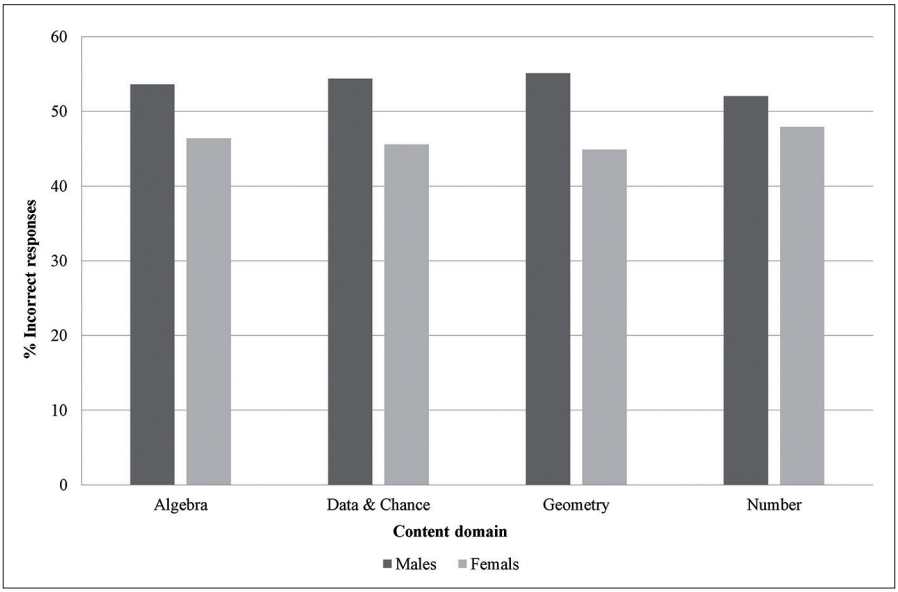


*Fig. 4 – Gender difference in Applying cognitive domains*

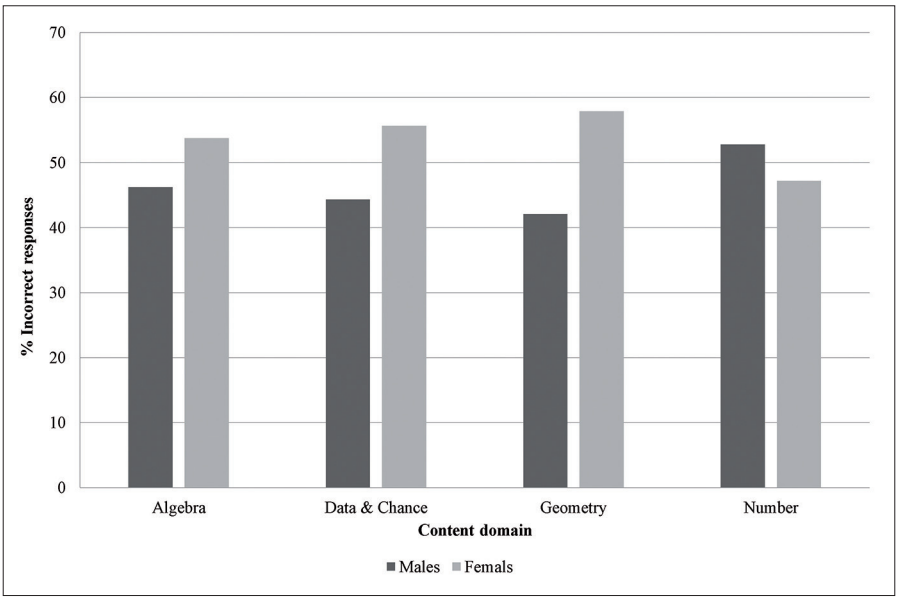


*Fig. 5 – Gender differences in Reasoning cognitive domains*



ISBN 9788835131557

The items considered in both content domains have a higher percentage of difficulty in the South and the South-Islands: *Applying* (of 100 students assigned this question, 89 students provided incorrect answers compared to 80% in the North and 85% in the Centre) and *Reasoning* (of 100 students assigned this question, 82 provided incorrect answers compared to 73% in the North and 74% in the Centre).

## 6. Concluding remarks

This study made it possible to define the test as an opportunity for both teachers and students. The identification of the errors made by students could become a learning opportunity to close the gap that generated them and thus could become a starting point for authentic teaching strategies to be implemented to overcome difficulties.

Following the reading and analysis of the single answers, the results of this study show that behind every single error in Mathematics there can be a variety of interacting factors: conceptual errors, such as the failure to consider the presence of the positive/negative sign or the failure to identify the fraction that has a higher value; errors characterized by a lack of understanding of the stimulus; or errors that can be linked to the student's expectation of how to complete the task, mainly characterized by the search for a possible model typical of a mathematical problem, even when the answer does not require the support of any type of calculation.

In this way, standardized tests can be transformed into a tool to support teaching, in what we might call "teaching with the test" (De Hoyos, Ganimian and Holland, 2017) as opposed to the undesirable practice of "teaching to the test".

Moreover, the Mathematics test is based on questions asked in Italian and therefore requires mother tongue knowledge and understanding. Incorrect or even vague answers can also be related to a difficulty in understanding and interpreting the question rather than a lack of mathematical skills.

Precisely in this regard, it would be useful to study the types of errors made by the students more deeply on an international basis, attempting to decipher, among the various participating countries, possible similarities in recurring errors and further investigating gender differences in the cognitive domains of *Applying* and *Reasoning*.

In this regard, we recall that Italy is one of the countries where marked differences between males and females in Mathematics are observed in all the different surveys and in all the school levels investigated. The specific



ISBN 9788835131557

knowledge of the different types of errors that the two genders commit in Mathematics tasks can be a valuable support for teaching, so that it is tailored to the cognitive features of each of the two genders.

Moreover, since only the mathematics test administered in the eighth grade was considered in this work, further research could compare these results with the results of the INVALSI national assessment since Mathematics is one of the subjects evaluated, together with the Italian and English languages, and the eighth grade is one of the grades involved in the national assessment.

## References


Allsopp D.H., Kyger M.M., Lovin L.H. (2007), *Teaching mathematics meaningfully*, Baltimore, Paul H. Brookes.

Braga M., Checchi D. (2010), "Sistemi scolastici regionali e capacità di sviluppo delle competenze. I divari dalle indagini PIRSL e PISA", *Italian Journal of Social Policy*, 3, pp. 1-25.

Cogan L.S., Schmidt W.H., Wiley D.E. (2001), "Who takes what math and in which track? Using TIMSS to characterize US students' eighth-grade mathematics learning opportunities", *Educational Evaluation and Policy Analysis*, 23, 4, pp. 323-341.

De Hoyos R., Ganimian A.J., Holland P.A. (2017), *Teaching with the test: experimental evidence on diagnostic feedback and capacity building for public schools in Argentina*, Policy Research Working Paper 8261, World Bank, Washington (DC).

Gardner H. (1991), *The Unschooled Mind: How children think and how schools should teach*, Basic Books, New York; tr. it. *Educare al comprendere. Stereotipi infantili e apprendimento scolastico*, Feltrinelli, Milano, 1993.

Herholdt R., Sapire I. (2014), "An error analysis in the early grades mathematics. A learning opportunity?", *South African Journal of Childhood Education*, 4, pp. 42-60.

INVALSI (2016), Rapporto Nazionale Indagini IEA TIMSS 2015, *La rilevazione IEA: i risultati degli studenti italiani nell'indagine internazionale TIMSS 2015*, retrieved on April 6, 2021, from: https://www.INVALSI.it/INVALSI/ri/timss2015/index.php?page=timss2015_it_05.

Ketterlin-Geller L.R., Yovanoff P. (2009), *Diagnostic assessments in mathematics to support instructional decision making*, retrieved on April 6, 2021, from: https://scholarworks.umass.edu/pare/vol14/iss1/16/.

Lester F.K., Kehle P.E. (2003), "From Problem Solving to Modeling: The Evolution of Thinking About Research on Complex Mathematical Activity", in R. Lesh, H.M. Doerr (*eds.*), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*, Lawrence Erlbaum Associates Publishers, Mahwah (NJ), pp. 501-517.




ISBN 9788835131557

Nesher P. (1987), “Towards an instructional theory: The role of student’s misconceptions”, *For the learning of mathematics*, 7, 3, pp. 33-40.
Priyani H.A., Ekawati R. (2018, January), “Error analysis of mathematical problems on TIMSS: A case of Indonesian secondary students”, in *IOP Conference Series: Materials Science and Engineering*, Vol. 296, No. 1, p. 012010, IOP Publishing.
Richards L., Morse J.M. (2007), *Fare ricerca qualitativa*, FrancoAngeli, Milano.
Russell M., Masters J. (2009), *Formative diagnostic assessment in Algebra and Geometry*, paper presented at the annual meeting of the American Education Research Association. San Diego, California.
Yang C.W., Sherman H., Murdick N. (2011), “Error pattern analysis of elementary students with limited english proficiency Investig.”, *Math. Learn.*, 4, pp. 50-67.
Zan R. (2007), *Problem solving. Difficoltà in matematica: Osservare, interpretare, intervenire*, Springer, Milano.



ISBN 9788835131557

# *3. Large Scale Assessment (LSA): a tool for mathematics education research*

by George Santi, Giorgio Bolondi, Federica Ferretti

Criticism against Large-Scale Assessment (LSA) by Mathematics educators developed because LSA has been interpreted as a behaviourist tool that looks at stimulus-response correlations, without unveiling the cognitive, emotional and social features behind the students' attitude towards the variety of items they are exposed to. Furthermore, Mathematics Education Research (MER) has firmly established paradigms: experimental designs, methodologies. These paradigms are mainly qualitative, thereby disregarding quantitative approaches. This approach to LSA has radically changed in the past years. LSA is an effective tool – based on a robust Mathematics education and statistics theoretical frameworks – to assess the learning of Mathematics of a whole system, with its educational, didactical, cultural-historical and political implications. There is a growing interest in broadening LSA's impact beyond the evaluation of school systems, allowing the use of materials from LSA in articulated research designs in MER. We refer to Theoretical Framework, Context-related information, Released Items, Global and Local Results, Micro-data etc. There are basically two reasons for introducing LSA in MER: 1) LSA is essential to take into account didactical macrophenomena that emerge from the complexity (in the sense of chaos theory) of teaching-learning processes, LSA allows us to highlight results from MER; 2) LSA brings into the research practice a methodology that can enhance the epistemological statute of Mathematics education, based on qualitative approaches. We introduced LSA in the research practice according to a mixed-method approach, intended as a self-contained methodological block, structured along the scheme: QUAL-QUAN-QUAL+QUAN. This methodological block is sustained by two legs: 1) theoretical framework appropriate for the Mathematical issue under study; 2) structured repository of LSA tools. The QUAL phase uses qualitative tools in a broad sense to single



ISBN 9788835131557

out the didactical variables and the research questions of the specific study. The QUAN phase is based on the use, according to the research questions and the didactical variables, of structured repositories. The QUAN+QUAL phase combines the quantitative data extracted from the repositories and the theoretical lens in order to answer the research questions, outline macro phenomena, confirm solid findings or highlight a new aspect regarding the learning of Mathematics. We discuss how several researches have been implemented according to that scheme, fulfilling three criteria: their results are coherent with results coming from previous researches; they highlight articulations of known results that had not been observed before; they point out new phenomena that deserve further studies in order to be explained. This shows that our approach can be considered a validated methodological model for intertwining LSA's materials and traditional paradigms in MER.

*Le critiche verso la Valutazione su Larga Scala (VLS) sono dovute al fatto che è stata intesa in senso comportamentista, come correlazione stimolo-risposta che non evidenza gli aspetti cognitivi, emotivi e sociali alla base delle risposte degli studenti. Inoltre, la Ricerca in Didattica della Matematica (RDM) dispone di paradigmi di ricerca solidi e radicati, prevalentemente qualitativi, che trascurano quelli quantitativi. L'atteggiamento nei confronti della VLS è cambiato radicalmente negli ultimi dieci anni. La VLS è uno strumento efficace, basato su solide cornici teoriche di didattica della Matematica e Statistica, per valutare l'apprendimento di sistema, con le sue implicazioni educative, didattiche, storico-culturali e politiche. Si assiste a un crescente interesse nell'estendere l'impatto della VLS oltre la valutazione dei sistemi scolastici, per usare i materiali della VLS in disegni di ricerca propri della RDM. Ci riferiamo alle cornici teoriche, alle informazioni di contesto, agli item rilasciati, ai risultati globali e locali, ai micro-dati ecc. Riteniamo che ci siano due motivi per introdurre la VLS nella RDM: la VLS è essenziale per interpretare i macro-fenomeni didattici che emergono dalla complessità (nel senso della teoria del caos) dei processi di insegnamento-apprendimento e consente di evidenziare i risultati che derivano dalla RDM; la VLS introduce nella ricerca una metodologia che può rafforzare lo statuto epistemologico della didattica della Matematica, basato su approcci qualitativi. Abbiamo introdotto la VLS nella pratica di ricerca utilizzando un metodo misto da considerarsi come un blocco metodologico auto-contenuto, strutturato secondo lo schema seguente: QUAL-QUAN-QUAL+QUAN. Il blocco metodologico è sostenuto da due gambe: una cornice teorica adeguata alla questione Matematica che si studia; un archivio strutturato degli strumenti della VLS. La fase QUAL usa strumenti qualitativi in un senso ampio per individuare le*



ISBN 9788835131557

*variabili didattiche e le domande di ricerca di uno studio specifico. La fase QUAN si basa sull'utilizzo, guidato dalle domande di ricerca e le variabili didattiche, di archivi strutturati. La fase QUAN+QUAL combina i dati quantitativi estratti dagli archivi con le lenti teoriche per rispondere alle domande di ricerca, delineare macrofenomeni, confermare risultati di ricerca consolidati o evidenziare un nuovo aspetto concernente l'apprendimento della Matematica. Discutiamo il modo in cui numerose ricerche sono state implementate secondo tale schema che deve soddisfare tre criteri: i risultati sono coerenti con risultati di ricerche precedenti; evidenziano articolazioni di risultati noti che non sono state osservate prima; indicano nuovi fenomeni che meritano studi ulteriori per essere spiegati. Tali criteri mostrano che il nostro approccio può essere considerato un modello metodologico validato per intrecciare i materiali della VLS con i paradigmi tradizionali della RDM.*

## 1. Introduction

In the past ten years, the National Assessment Institute (INVALSI) has administered to Italian students Italian Language and Mathematics tests. There has been, and to a certain extent there still is, criticism against large scale assessment (LSA) on the part of the mathematics education community. Mainly because LSA has been interpreted as a behaviourist tool that looks at stimulus-response correlations, without unveiling the cognitive, emotional and social processes beyond the students' attitude towards the variety of items they are exposed to in the mathematics tests.

This approach to LSA has radically changed in the past years. LSA is an effective tool – based on a robust mathematics education and statistics theoretical frameworks – to assess the learning of mathematics of the whole national school system, with its educational, didactical, cultural-historical and political implications.

There is a growing interest in broadening LSA beyond the evaluation of school systems to mathematics education research as a new methodological tool (de Lange, 2007; Meinck, Neuschmidt and Taneva, 2017); we refer to Theoretical Framework, Context-related information, Release Items, Global and Local Results, Micro-data etc. Furthermore, mathematics education research has traditionally and firmly established paradigms: experimental designs, methodologies. These paradigms have been mainly qualitative (Hart *et al.*, 2009) disregarding the quantitative ones. The attention of the mathematics education community towards LSA allows for the introduction of quantitative paradigms to enlarge the range of research methodologies.



ISBN 9788835131557

We believe there are basically three reasons for introducing LSA in mathematics education:

1) LSA is essential to take into account didactical macrophenomena that emerge from the complexity (in the sense of Chaos Theory) of teaching-learning processes at the level of a school system;
2) LSA allows us to highlight research results in mathematics education;
3) LSA brings into the research practice a new methodology that can enhance the epistemological statute of mathematics education, based on qualitative approaches.

The aim of the present chapter is to investigate the role of INVALSI Large Scale Assessment in fostering the development of mathematics education research paradigms, both theoretical and methodological.

In Section 2, we show the theoretical tenets of INVALSI Large Scale Assessment and their connections with mathematics education theoretical perspective. In Section 3, we present a new methodological block that intertwines qualitative methodologies with Large Scale quantitative ones. In Section 4, we show two implementations of such a methodological block regarding the learning of high school algebra, as a macrophenomenon emerging from the Italian mathematical school system. In Section 5 we suggest some conclusion that can be drawn from our study.

## 2. INVALSI theoretical perspective

One of the main factors that makes the INVALSI tests valid from an educational point of view, is that they are in line with ministerial regulations (National Guidelines for the first cycle of education and Kindergarten schools; National Guidelines for High Schools, and Guidelines for technical/professional institutes, for secondary education), as well as with the main results of national/international research findings in mathematics education. This allows the collection of samples of assessment tests and the analysis of results, focusing on knowledge and skills required by scholastic curricula, and often investigating difficulties highlighted by the literature.

We recall some of the main theoretical strands in mathematics education that inform the construction of the INVALSI tests:

– the Theory of Didactical Situations (Brousseau, 2002) and the effects of the didactical contract (D'Amore, 1999);
– the Triangle of Chevallard and the Didactical Transposition (Chevallard and Joshua, 1982);



ISBN 9788835131557

– the role of semiotics in mathematical thinking and learning in its structural-functional approach (Duval, 1995) and the sociocultural ones (Arzarello, 2006; Radford, 2010; Bartolini-Bussi and Mariotti, 2008; Godino, Batanero and Font, 2007);
– the theory of Obstacles (Brousseau, 1983);
– Fischbein's (1993) Theory of Figural Concepts;
– the role of mental images and mental models in mathematical learning and the emergence of misconceptions (D'Amore, 1999; Sbaragli, 2005);
– argumentation and proof in Mathematics (Duval, 1996; Hanna and de Villiers, 2012).

Another important contribution afforded by mathematics education is a range of solid findings that allow INVALSI to interpret quantitative data. Large scale quantitative data provide interesting results from a statistical point of view, but we do not have access to the cognitive processes that underlie the student's answers to the tests. The combination of the quantitative statistical information with the results of mathematics education research is a powerful tool both for LSA and educational research.

If, on the one hand, INVALSI profits from mathematics education theoretical and experimental research, on the other hand the opposite is also true. LSA offers the scientific community data that require new interpretations and further developments of acknowledged theoretical perspectives.

"Solid finding" is a category of Mathematics Education (EMS, 2011; Bosch *et al.*, 2017). These findings are validated through shared research paradigms and methodologies, with a prevalence of a qualitative approach (Hart *et al.*, 2009). Having a quantification of the magnitude of the phenomena highlighted by the research which might help teachers when facing their specific teaching-learning situations.

Of course, what is needed is to integrate suitable theoretical lenses and with information on the context.

An emblematic example is given by Ferretti and Bolondi (2019). Their research shows how, from LSA data, a new effect of the didactical contract emerges.

Another interesting contribution that LSA can bring to mathematics education research is to fulfill the need of a systemic approach – in the sense of Chaos Theory – to mathematics teaching and learning. Mathematics education research paradigms are mainly qualitative and involve case studies, longitudinal studies, low number of students. This approach is extremely effective to investigate mathematical cognitive and learning processes. Nevertheless, they cannot encompass the generality and complexity of teaching and learning at the level of a school system. LSA brings to the fore new di-



ISBN 9788835131557

dactical phenomena (Ferretti and Bolondi, 2019) and requires to re-interpret solid findings from a systemic stance. The systemic approach entailed by LSA on the one hand broadens the range of action theoretical perspectives in mathematics education on the other hand it contributes to the development of new paradigms.

Therefore, LSA quantitative data can serve as a further epistemological foundation of mathematics education by contributing to the development of the field with a new methodological approach.

## 3. A new methodological block

In the previous section, we have shown the interplay between LSA and mathematics education research from a theoretical and epistemological point of view. In the present, section we show the outcome of such an interplay from a methodological point of view. In fact, as pointed out by Radford (2008) a theory is a triadic structure that consists of a system of principles, a methodology and a template of research questions. The true underpinning of the triad is the system of principles in which both the methodology and the research questions are embedded. There is a coherent relation between the structure of the system of principle, the methodology and the set of possible research questions that stem from a theory. The broader theoretical "space" that emerges from the dialogue between LSA and mathematics education research has consistently developed into a new methodology in educational research.

We introduced a new research methodology for mathematics education based on the insertion of LSA theoretical and experimental paradigms in research practice.

We have outlined a new self-contained methodological block that encompasses qualitative and quantitative elements, structured along the following scheme introduced by Johnson and Onwuegbuzie (2004):

QUAL → QUAN → QUAL+QUAN.

This methodological block is sustained by two legs:

1) a theoretical framework appropriate for the mathematical issue under study;
2) a structured repository of LSA tools (Bolondi, Ferretti and Gambini, 2017; Ferretti, Gambini and Santi, 2020).

The theoretical framework is related not only to the mathematical content involved in the study but also to the features of the complex system from which the macro-phenomena emerge.

With regard to the repository, a group of mathematics education researchers (ForMATH Project), in collaboration with computer scientists, on behalf



ISBN 9788835131557

of INVALSI introduced in 2014 a new tool that teachers could use in order to bring the standardized assessment into their school practice and professional development. We are referring to GESTINV, a database with structured information regarding Italian standardized assessment that contains 1,718 test items, spanning 10 years of INVALSI activity. The database has been devised both for Italian and Mathematics. Entering the Mathematics section, you can search according to: ministerial regulations; keywords (there are about 200 keywords that identify the main topic for each item); full text of an item by typing keywords; the INVALSI theoretical framework; national rates of correct/incorrect/invalid answers; types of test questions (multiple choice, open questions, etc.): guided cross search (with and/or logical connectors) involving all the parameters mentioned above.

We now enter into the details of the phases that constitute our methodological block.

Our research method is based on a quantitative methodology, driven by semiotic theoretical perspective, that utilizes the results deriving from large scale assessment.

### *3.1. The QUAL phase*

This phase requires to pinpoint the research focus and the suitable and effective theoretical lenses. A clear research focus along with is theoretical framework allows us to identify the research questions and the didactical variables. They will be of extreme importance in the interpretation of the data and in the selection of the macro-phenomena.

The QUAL phase uses qualitative tools in a broad sense not only to single out the research questions and the didactical variables of a specific study but also to provide the features of the complex system in which the phenomena are embedded. The QUAL phase resorts to several tool such as theoretical perspectives, interviews, discussion groups, group activity, on-the-ground observations, etc.

### *3.2. The QUAN phase*

This phase is based on the implementation of GESTINV (Ferretti, Gambini and Santi, 2020) exploiting its rich resources in terms of available items of the INVALSI tests indexed according to the National Guidelines, the results from the statistical point of view, the mathematical content, the key



ISBN 9788835131557

words, the percentage of correct, wrong and invalid answers and the other characteristics mentioned above.

Ferretti, Giberti and Lemmo (2018) provide significative examples of the use of GESTINV in mathematics education research. GESTINV allows us to carry out a quantitative analysis based on the INVALSI tests pertaining the research focus, the research questions and the didactical variables. We point out that the selection of the most significative items using GESTINV is strongly driven by the QUAL phase not only with regard to the research questions and the didactical variables, but also to other systemic characteristics that include contextual educational and socio-economic information. The functions of GESTINV provide items that match the research needs in terms of cognitive processes, mathematical content and learning objectives underpinning the research questions of the investigation.

### *3.3. The QUAN+QUAL phase*

The QUAN+QUAL phase combines the quantitative data extracted from GESTINV and the theoretical lens in order to answer the research questions, outline macro phenomena, confirm solid findings or highlight a new aspect regarding the learning of mathematics that require to broaden a theoretical perspective or network existing ones (Prediger, Bikner-Ahbahs and Arzarello, 2008).

The qualitative and quantitative features are used with different nuances with respect to the previous phases. The qualitative aspect refers to the implementation or broadening of theoretical perspectives for the interpretation of data, the answer to the research questions and the outlining of macrophenomena. The quantitative aspect refers to the statistical information provided by GESTINV after the QUAN phase in terms of the characteristic curves, distractor plots, ITN, etc. This stage entangles the qualitative variables with the quantitative ones in order to implement a strong and effective tool for the interpretation of macro-phenomena emerging from the Italian mathematics educational system conceived as a complex system in terms of Chaos Theory.

In the light of the results emerging from the QUAN Phase, the QUAL+QUAN one could require not only the broadening of the theoretical framework but also the consistent reformulation of the research questions.



ISBN 9788835131557

## 4. Two examples

We present two studies picked out from a research program concerning the learning of Algebra in Italian high schools, grade 10 students, conducted by the authors. Both studies are a significative implementation of the methodological block that we described in Section 3. Space limitations do not allow us to display the whole research but we will only go through the basic key points.

### *4.1. Powers in grade 10*

Our first case study for validating our methodological block focused on syntactic aspects because INVALSI tests show that Italian high school students have severe difficulties in handling the meaning of algebraic formalism when dealing with powers. This case will be presented in more detail in a forthcoming paper.

#### *4.1.1. QUAL phase*

The research focused, within the Fregean approach to meaning, on the relation between semiotic expressions, sense and denotation (Arzarello, Bazzini and Chiappini, 2001). The researchers' interest was on the students' ability to face the special ontological and epistemological character of mathematical objects/concepts, that is, their intrinsic inaccessibility due to their ideal nature. Therefore, students identify the signifier with the signified, the sense of the algebraic expression (the signifier) with its denotation (the signified). Furthermore, exposed to different algebraic expressions of the same object/concept they identify each expression with a different object; that is, different signifiers of the same object/concept with different signified. We term this phenomenon as a change of meaning due to treatment algebraic transformations (D'Amore, 2007).

We framed the learning of powers within the structural and functional approach to semiotic introduced by Duval (1995; 1996). He highlights a specific cognitive functioning in mathematics, due to the special epistemological nature of its objects that do not allow ostensive references. Thinking and learning in mathematics is identified with the coordination of semiotic registers via treatment and conversion. Treatment is a semiotic transformation from a representation into another within the same semiotic system and



ISBN 9788835131557

conversion is a semiotic transformation from a representation in one semiotic system into another representation in another semiotic system.

We identified the following research questions:

– Q1: What precise information, regarding powers, can we acquire from a research that implements GESTINV?
– Q2: Is it possible to collect information that is coherent with solid research findings?

### *4.1.2. QUAN phase*

GESTINV allowed us to carry out a quantitative analysis based on the INVALSI grade ten mathematics items, selecting the ones with lower scores. Among these, we noticed that the management of powers that required treatment operations yielded the worst results.

We present only one of the items selected from our data.

The expression $a^{43} + a^{44}$ is equal to:

A. ☐ $a^{44 \cdot 43}$

B. ☐ $a^{43} \cdot (a+1)$

C. ☐ $a^{87}$

D. ☐ $2a^{87}$

*Fig. 1 – Task in Mathematics grade 10 INVALSI test 2015*

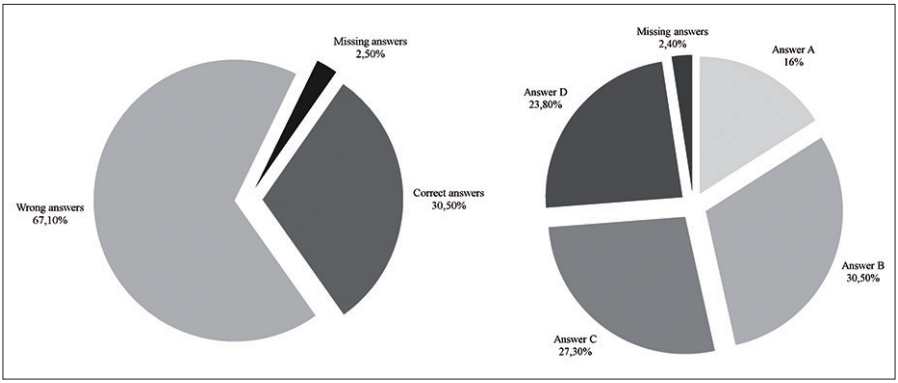


*Fig. 2 – Results referred to the task in Mathematics grade 10 INVALSI test 2015*



ISBN 9788835131557

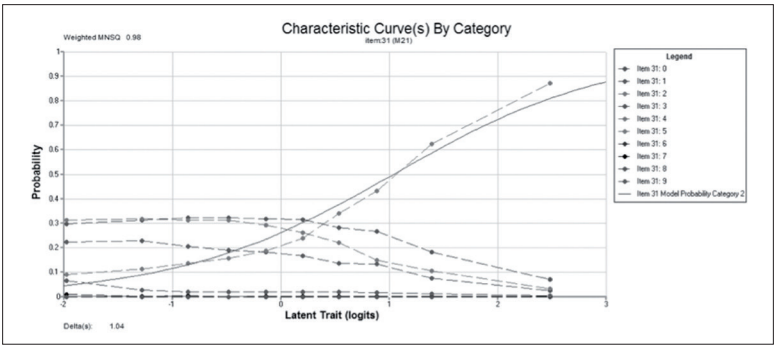


*Fig. 3 – Characteristic Curve referred to the task in Mathematics grade 10 INVALSI test 2015*

### *4.1.3. QUAN+QUAL phase*

In 2015 almost 550,000 grade 10 students performed this INVALSI test, and the national results referred to a sample of 4,8440 students. As we can see in the following graphs, only a third of Italian students provided the correct answer; among the incorrect options, the most chosen option is C. In option C the exponents of the two powers in the text are summed. Again, this protocol shows a loss of meaning, due to a treatment, when the student goes from the original $a^{43}+a^{44}$ to $a^{87}$. The expression $a^n+a^m$ puzzles the student who is not able to frame it appropriately in the context of powers, thereby he resorts to the well-known identity $a^n*a^m=a^{(n+m)}$ which leads to a loss of the original meaning. This result suggests that factoring out the GCF is meaningless to most students, despite the thorough practice in terms of treatment transformations they are exposed to. Meaningless in the sense that they confuse the algebraic representations (signifiers) with the mathematical object (signified) and they are not able to establish the correct semiotic reference to the mathematical object.

The Characteristic Curve (Figure 3), shows that, among the incorrect options, Option C is the most chosen at all levels of competence.

The QUAN+QUAL analysis allows us to answer to the research questions:

A1) at a coarse-grained level, GESTINV highlights rooted difficulties in students facing treatments regarding powers. This is a quantitatively relevant macro phenomenon. INVALSI results highlight that it persists with the same features across time;



ISBN 9788835131557

A2) at a fine-grained level, GESTINV provides data that are coherent with solid research findings. In particular, it unveils at a quantitative level the phenomenon of change/loss of meaning due to treatment semiotic transformations. Several studies (D'Amore, 2007; D'Amore and Fandiño Pinilla, 2007; Santi, 2011) have shown that at all school levels, including prospective teachers, also treatment bewilders students who experience a loss or a change of meaning in treatment transformations. The loss and/or change of meaning due to treatment transformations implies that mathematical cognition in general and in particular the algebraic one cannot be reduced to a complex transformation of signs. Meaning is beyond the mere relation sign-object and it is necessary to take into account other basic features that characterize sense-making processes in mathematics. In particular, the present study reveals a different instance of loss of meaning with respect to the previous research, mentioned above. In fact, students identify different algebraic expressions, which refer to different objects/concepts, with the original sum of powers represented by the original expression in the item. In a more general sense, the incorrect relation between expression, sense and denotation confirms another important solid finding, known as Duval's (1995) cognitive paradox that impels students to identify semiotic representations with the mathematical object. Moreover, data easily available in GESTINV show that this phenomenon also affects students with medium-high levels of competences. Thus, teacher's didactical awareness is not only aimed at helping weak students but also the so-called stronger ones, devising an effective didactical transposition and didactical engineering that encompass the complexity of mathematical thinking and learning;

A3) GESTINV is an effective tool that entangles quantitative and qualitative research methodologies. As regards the quantitative approaches, they are based on a statistically significant population. It allows us to provide a quantitative validation of theoretical results, confirmed at a qualitative level. Furthermore, the characteristic curves are a powerful tool to intertwine quantitative and qualitative analyses.

### *4.2. Inequalities in grade 10*

The second study belongs to the same research program mentioned in the previous section. Our focus was on the learning of inequalities related to the semantic control when dealing with the treatment of algebraic representations. The interest in inequalities was driven both by research in mathematics



ISBN 9788835131557

education (Linchevsky and Sfard, 1992; Tsamir, Almog and Tirosh, 1998) and by educational issues at the level of the Italian school system.

### *4.2.1. QUAL Phase*

The aim of the study was to test the students' acquaintance with treatment and conversions. High school algebra practice in Italy requires students to solve inequalities with different strategies that involve number processes, powers with positive exponents, cartesian geometry, control of the truth value in propositional logic; all these strategies basically pivot around a strong syntactic control of algebraic calculations. From a semiotic point of view, the aforementioned strategies have their counterparts in terms of semiotic systems and treatment/conversion operations. The learning of algebra therefore requires a strong semiotic and theoretic control (Arzarello and Sabena, 2011) that can clash with Duval's cognitive paradox (Duval, 1995) and the students struggle in dealing with Frege's triad expression-sense-denotation (Arzarello, Bazzini and Chiappini, 2001).

We decided to frame our study within Duval's (1995) structural and functional approach resorting to the notions of semiotic system, choice of distinctive traits of semiotic representations, treatment and conversion.

We outlined the following research questions:

– How do the structural and functional potentials of the algebraic symbolic language networked with other semiotic systems allow the students to deal with inequalities?
– How do students solve inequalities resorting to meaning in terms of the referential relation between signifier-signified and in a more general understanding in terms of Frege's triangle expression-sense-denotation?

### *4.2.2. QUAN phase*

GESTINV allowed us to carry out a quantitative analysis based on the INVALSI grade 10 algebra items concerning treatment operations with inequalities. From the results of our research using GESTINV it turned out that the items with lower scores of correct answers were characterized by the fact that they could be solved having recourse to intuitive thinking or, in case of pseudo-structural students, by performing standard calculation to arrive at the solution.

We present only one of the items selected from our data that we reckon as particularly significative.



ISBN 9788835131557

In the set of real number the inequality $x^2 + 1 \geq 0$ is verified

A. ☐ only for $x \geq 0$

B. ☐ only for $x \geq -1$

C. ☐ for all $x$

D. ☐ for none $x$

*Fig. 4 – Task D02, INVALSI Mathematics test 2015, grade 10*

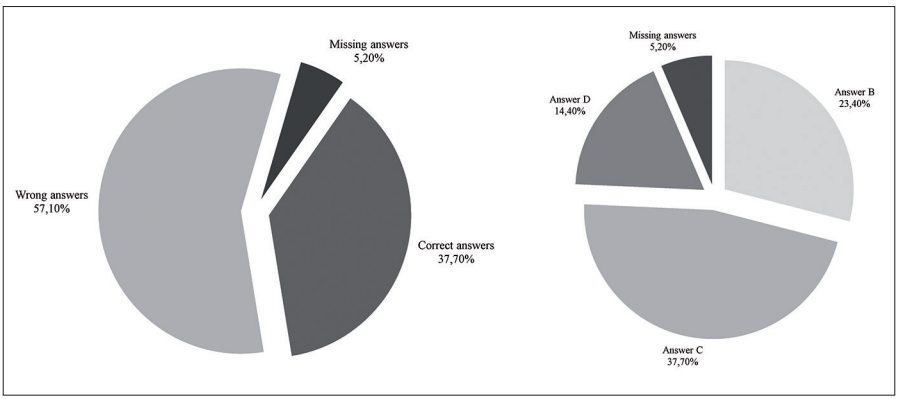


*Fig. 5 – Percentage of answers at national level, Task D02, INVALSI Mathematics test 2015, grade 10*

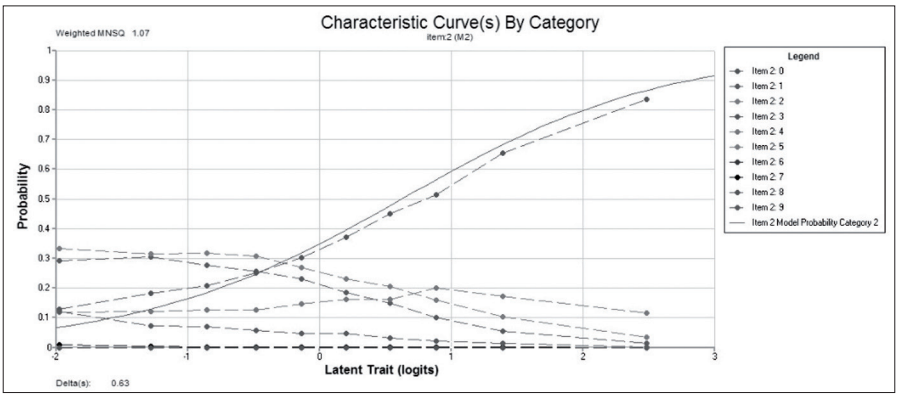


*Fig. 6 – Characteristic Curve, Task D02, INVALSI Mathematics test 2015, grade 10*



ISBN 9788835131557

### *4.2.3. QUAN+QUAL Phase*

This item addressed 550.000 grade 10 Italian students and the results refer to a sample of 27.000 students.

The correct answer is chosen by less than 40% of the students (Figure 6).

### *4.2.4. QUAN+QUAL Phase*

This item addressed 550.000 grade 10 Italian students and the results refer to a sample of 27.000 students.

The correct answer is chosen by less than 40% of the students (Figure 11). Distractor D, although the less chosen, scores a remarkable 14.4%, as predicted by the INVALSI guidelines, since the binomial is never equal to zero. Distractors A (19.3%) and B (23.4%) are the two most chosen distractors that altogether score more than 40%. As we can infer from the characteristic curve (Figure 6), distractors A and B are the most chosen up to high levels of competences and both are preferred to the correct answer up to the $4^{th}$ percentile.

As it is with other items emerging from the present research this result was completely unexpected. According to the INVALSI (2015) guidelines we expected the following strategies:

1) resorting to number strategies, we expected that students recognized that the sum of two squares is a non-negative number;
2) they could have analysed the function $f(x) = x^2+1$ and notice that its graph is always above the x-axis, so for all x, $x^2+1 \geq 0$;
3) they could have applied the general techniques for solving second degree equations. Observing that $\Delta<0$ and that the coefficient of $x^2$ is a real positive number, they could have concluded that $x^2+1 \geq 0$ for all $x \in R$. The $\geq$ could be an issue for those students who do not recognize that positive numbers can be greater or equal to 0. Freudenthal outlined the difference between telling the truth and telling all the truth. In mathematics we are satisfied with the truth and we accept that positive numbers are $\geq 0$.

The strategies mentioned above require a good control of the basic semiotic functions, in particular treatment and/or conversion. Strategies 1. and 2. require a network of treatments and conversions with a very high demand in the students' semiotic and theoretic control. If they were the only available strategies, we would have expected such low scores for the correct option. The choice of distractor D is justified by the last comment of strategy 3.

Our expectation was that students with a very weak semantic and theoretic control would at least have carried out the standard calculations de-



ISBN 9788835131557

scribed in strategy 3. that, in the worst cases, do not require any semantic control. The calculations are a series of treatments that are controlled by the transformation rules of the algebraic semiotic system. Furthermore, they are strongly trained in Italian algebra school practice, at a systemic level. Given the low percentage of correct answers, this was not a viable option for most students.

We testify an unforseen result that emerged as a macrophenomenon from the Italian Educational System conceived, as a chaotic system. We draw the attention of the reader on the twofold dimension of this phenomenon, quantitative and qualitative. The percentage of correct answer is around 35%, which is statistically significant and allows us to regard it as a macrophenomenon.

Our structural and functional semiotic approach is seemingly inappropriate to predict and fully understand this phenomenon. We remark that the solution of the inequality is not necessarily accountable for in terms of algebra, but an embodied and intuitive reasoning, in the domain of numbers, would have easily led to the correct answer of the item. This is an alarm bell that this question hits something very deep regarding students' personal meaning of inequalities that calls for a more effective theoretical interpretative lens.

The instance that students did not resort to mere mechanic calculations in the algebraic language tells that they did not even overcome their puzzlement towards the item by resorting to rituals (Lavie, Steiner and Sfard, 2019) nor to the clauses of the didactical contract (D'Amore, 1999). This result is extremely important in pinpointing the effectiveness of the INVALSI methodology. The Context of the assessment setting, the multiple-choice item and structure of the distractors have unmasked a cognitive behaviour that would not have appeared in a standard classroom setting within Chevallard's triangle Knowledge-student-teacher (Chevallard and Joshua, 1982).

The need to appropriately outline this emerging macrophenomenon, prompted the construction of a new conceptual framework resulting from the networking of different theoretical perspectives (Prediger, Bikner-Ahbahs and Arzarello, 2008). We consequently reformulated our research questions, embedded in this new conceptual framework. Space limitations do not allow us to describe this new theoretical framework that will be the focus of future publication.



ISBN 9788835131557

## 5. Concluding remarks

LSA is a fully-fledged practice spread all over the world, performed both by single nations and international organizations such as OCSE-PISA and TIMMS. Its range overarches political, sociological, educational and epistemological issues. The potentials and the strength LSA cannot remain confined to the ranking of students, schools and nations. We have to develop a new dialogue between LSA and mathematics education in order to fully acknowledge the enormous potentials and the educational aims of both practices. Their fruitful encounter is possible on the basis effective theoretical tools to interpret the quantitative and cutting-edge research methodologies to acquire interesting data and interpret the macrophenomena that emerge from the complexity of educational systems.

LSA can truly improve the teaching and learning of mathematics only if it is able to give refined, culturally wide-ranging and operational information to policy makers, teacher training programs, curriculum developers, principals and teachers. This is not possible without a profound connection between LSA and mathematics education research.

The present study addresses the methodological features regarding the intertwining of LSA and mathematics education research. We introduced LSA in the research practice according to a mixed-method approach, intended as a self-contained methodological block, structured along the scheme:

QUAL-QUAN-QUAL+QUAN

This methodological block is sustained by two legs:

1) a theoretical framework appropriate for the mathematical issue under study;
2) a structured repository of LSA tools.

The QUAL phase uses qualitative tools in a broad sense to single out the didactical variables and the research questions of the specific study. The QUAN phase is based on the use, according to the research questions and the didactical variables, of structured repositories. The QUAN+QUAL phase combines the quantitative data extracted from the repositories and the theoretical lens in order to answer the research questions, outline macro phenomena, confirm solid findings or highlight a new aspect regarding the learning of mathematics.

We discussed two researches that have been implemented according to that scheme, fulfilling three criteria:

– their results are coherent with results coming from previous researches;
– they highlight articulations of known results that had not been observed before;
– they point out new phenomena that deserve further studies in order to be explained.



ISBN 9788835131557

This shows that our approach can be considered a validated methodological model for intertwining LSA theoretical framework and methodology with traditional paradigms in mathematics education research.

## References


Arzarello A., Bazzini L., Chiappini G. (2001), "A model for analyzing algebraic process of thinking", in R. Sutherland, T. Rojano, A. Bell (*eds.*), *Perspectives on school algebra*, Kluwer Academic Publisher, Dordrecht (NL), pp. 61-82.

Arzarello F. (2006), "Semiosis as a multimodal process", *RELIME*, Numero Especial, pp. 267-299.

Arzarello F., Sabena C. (2011), "Semiotic and theoretic control in argumentation and proof activities", *Educational Studies in Mathematics*, 72, 2-3, pp.189-206.

Bartolini Bussi M.G., Mariotti M.A. (2008), "Semiotic mediation in the mathematics classroom", in L. English (*eds.*), *Handbook of international research in mathematics education*, Routledge, New York/London, 2nd ed., pp. 746-783.

Bolondi G., Ferretti F., Gambini A. (2017), "Il database GESTINV delle prove standardizzate INVALSI: uno strumento per la ricerca", in P. Falzetti (a cura di), *I dati INVALSI: uno strumento per la ricerca*, FrancoAngeli, Milano, pp. 33-42.

Bosch M., Dreyfus T., Primi C., Shiel G. (2017), *Solid findings in mathematics education: What are they and what are they good for?*, CERME 10, February, Dublin, Ireland, retrieved on April 6, 2021, from hal-Archives Ouvertes, HAL Id: hal-01849607, https://hal.archives-ouvertes.fr/hal-01849607.

Brousseau G. (1983), "Les obstacle épistémologiques et les problèmes en mathémathiques", *Recherches en didactique des mathématiques*, 4, 2, pp. 165-198.

Brousseau G. (2002), *Theory of didactical situations in mathematics: Didactique des mathématiques, 1970-1990*, Kluwer Academic Publishers, New York/Boston/Dordrecht/London/Moscow.

Chevallard Y., Joshua M.A. (1982), "Un exemple d'analyse de la transposition didactique: la notion de distance", *Recherches en didactique des mathématiques*, 3, 1, pp. 159-239.

D'Amore B. (1999), *Elementi di didattica della matematica*, Pitagora, Bologna.

D'Amore B. (2007), "Mathematical objects and sense: how semiotic transformations change the sense of mathematical objects", *Acta Didactica Universitatis Comenianae*, 7, pp. 23-45.

D'Amore B., Fandiño Pinilla M.I. (2007), *Change of the meaning of mathematical objects due to the passage between their different representations. How other disciplines can be useful to the analysis of this phenomenon. Rome, Symposium on the occasion of the 100th anniversary of ICMI, March 2008. WG5: The evolution of theoretical framework in mathematics education, organizers: Gilah Leder and Luis Radford*, retrieved on April 6, 2021, from: www.unige.ch/math/EnsMath/Rome2008.




ISBN 9788835131557

De Lange J. (2007), "Large-Scale Assessment and Mathematics Education", in F.K. Lester Jr. (*ed.*), *Second Handbook of Research on Mathematics Teaching and Learning*, National Council of Teachers of Mathematics (NCTM), Charlotte (NC), pp. 1111-1142.

Duval R. (1995), *Sémiois et pensée humaine. Registres sémiotiques et apprentissages intellectuels*, Peter Lang, Berne.

Duval R. (1996), "Argomentare, dimostrare, spiegare: continuità o rottura cognitiva?", *La matematica e la sua didattica*, 2, pp. 130-150.

EMS (2011), "'Solid Findings' in Mathematics Education", *Newsletter of the European Mathematical Society*, 81, pp. 46-49.

Ferretti F., Bolondi G. (2019), "This cannot be the result! The didactic phenomenon 'the age of the earth'", *International Journal of Mathematical Education in Science and Technology*, 52, 2, pp. 194-207.

Ferretti F., Gambini A., Santi G. (2020), *The GESTINV Database: A Tool for Enhancing Teachers Professional Development within a Community of Inquiry, in H. Borko, D. Potari (eds.), Pre-Proceedings of the Twenty-fifth ICMI Study School Teachers of Mathematics working and learning in collaborative groups*, University of Lisbon, Lisbon, pp.621-628.

Ferretti F., Giberti C., Lemmo A. (2018), "The Didactic Contract to Interpret Some Statistical Evidence in Mathematics Standardized Assessment Tests", *EURASIA Journal of Mathematics, Science and Technology Education*, 14, 7, pp. 2895-2906.

Godino J.B., Batanero C., Font V. (2007), "The ontosemiotic approach to reseach in mathematics education", *ZDM Mathematics Education*, 39, pp. 127-135.

Hanne G., de Villiers M. (2012), *Proof and proving in mathematics education. The 19th ICMI Study*, Springer, Dodrecht (NL).

Hart L.C., Smith S.Z., Swars S.L., Smith M.E. (2009), "An Examination of Research Methods in Mathematics Education (1995-2005)", *Journal of Mixed Methods Research*, 3, 1, pp. 26-41.

Lavie I., Steiner A., Sfard A. (2019), "Routines we live by: From ritual to exploration", *Educational Studies in Mathematics*, 101, 2, pp. 153-176.

Linchevsky L., Sfard A. (1992), "Rules without reasons as processes without objects – the case of equations and inequalities", in F. Furinghetti (*ed.*), *Proceedings of the 15th Conference of the International Group for the psychology of Mathematics Education*, PME, Assisi, vol. 2, pp. 317-324.

Meinck S., Neuschmidt O., Taneva M. (2017), "Workshop Theme: 'Use of Educational Large-Scale Assessment Data for Research on Mathematics Didactics'", in G. Kaiser (*ed.*), *Proceedings of the 13th International Congress on Mathematical Education*, ICME-13 Monographs, Springer, Cham.

Prediger S., Bikner-Ahsbahs A., Arzarello F. (2008), "Networking strategies and methods for connecting theoretical approaches: first steps towards a conceptual framework", *ZDM Mathematics Education*, 40, pp. 165-178.

Radford L. (2017), "Mathematics education theories: The question of their growth, connectivity, and affinity", *La matematica e la sua didattica*, 25, 2, pp. 217-228.



ISBN 9788835131557

Santi G. (2011), “Objectification and semiotic function”, *Educational Studies in Mathematics*, 66, 77, pp. 285-311.
Sbaragli S. (2005), “Misconcezioni ‘inevitabili’ e misconcezioni ‘evitabili’”, *La matematica e la sua didattica*, 1, pp. 57-71.
Tsamir P., Almog N., Tirosh D. (1998), “Students’ solutions to inequalities”, in A. Olivier, K. Newstead (*eds.*), *Proceedings of the 22nd Conference of the International Group for the Psychology of Mathematics Education*, PME Program Committee, Stellenbosch, vol. 4, pp. 129-136.



ISBN 9788835131557

# *4. Assessment of differential item functioning: first comparisons on INVALSI test and some policy implications*

by Simone Del Sarto, Michela Gnaldi

Differential Item Functioning (DIF for short) is a bias of a test item, which occurs whenever the response probability to that item differs between groups of examinees with the same ability level (e.g., groups according to gender, geographic location, etc.). It is therefore important to verify, within a questionnaire, the presence of items affected by DIF, in order to avoid potential drawbacks in the validity as regards the single items and the test as a whole. In the literature several statistical methods have been proposed for identifying DIF, generally distinguished according to the framework they are based on (Classical Test Theory or Item Response Theory). In this work, we show a comparison between some approaches proposed for DIF detection, by applying them to data coming from an INVALSI test, in particular highlighting the agreement between them in detecting the item with DIF. Results show that DIF detection methods assuming unidimensionality of the latent trait essentially identify the same items with DIF. However, by introducing a multidimensional approach – which supposes several, potentially correlated, latent traits, affecting the item response process – we obtain a substantial disagreement with respect to unidimensional DIF detection methods. This situation is potentially misleading, as it leads to non-univocal considerations for certain items, which appear affected by DIF under a method and not affected by DIF under another. Thus, during a validation phase of a test and its items, it is essential to jointly carry out DIF detection through different approaches and perform a test dimensionality assessment, in order to evaluate which DIF detection approach (unidimensional or multidimensional) is the most suitable for the data at issue.

*Il* Differential Item Functioning *(DIF in breve) è una distorsione presente in una particolare domanda di un test che si manifesta ogniqualvolta la probabilità di risposta a quella domanda differisce tra gruppi di esaminandi*



ISBN 9788835131557

*con lo stesso livello di abilità (per esempio, gruppi costituiti in base a genere, area geografica ecc.). È dunque molto importante verificare se sussistano all'interno di un questionario item affetti da DIF e in quale proporzione, per evitare una possibile riduzione di validità degli item e del test nel suo complesso. In letteratura esistono vari metodi statistici per identificare il DIF, generalmente distinti a seconda del contesto su cui sono basati (teoria classica del test o Item Response Theory). In questo lavoro presentiamo un confronto tra alcuni approcci proposti per identificare il DIF, applicandoli ai dati relativi ad un test INVALSI, evidenziando in particolare la concordanza tra di essi nel rilevare gli item affetti da DIF. I risultati mostrano come i metodi che assumono unidimensionalità del tratto latente rilevino sostanzialmente gli stessi item affetti da DIF. Introducendo poi un approccio multidimensionale – che presuppone cioè più tratti latenti potenzialmente correlati che influenzano la risposta all'item – otteniamo un considerevole numero di item per cui i risultati non concordano con l'approccio unidimensionale. Questa situazione è potenzialmente fuorviante, in quanto porta a considerazioni non univoche per alcuni item. È importante quindi, in fase di validazione del test e dei suoi quesiti, utilizzare congiuntamente diversi approcci per identificare il DIF, così come è fondamentale effettuare un controllo della dimensionalità del test, in modo tale da poter valutare quale approccio di DIF detection (unidimensionale o multidimensionale) sia più opportuno data la struttura di dimensionalità del test.*

## 1. Introduction

In this work we show a comparison between some approaches proposed detecting Differential Item Functioning (DIF), by applying them to data coming from the INVALSI Mathematics test administered to Italian pupils of primary schools in 2019 with the main aim to verify the degree of agreement between DIF detection methods.

DIF, also known as item bias, occurs when subjects from different groups, for instance clustered on the basis of gender or geographic area and with the same level of the latent trait, have a different probability of giving a certain response to a given item. Tests containing such items may have a reduced validity for between-group comparisons, because their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen *et al.*, 1988). Therefore, detection of DIF is useful to validate a questionnaire and it may provide aid for interpreting the psychological process underlying group differences in answering to the given items.



ISBN 9788835131557

There are various methods to identify items affected by DIF. They differentiate each other on the basis of their development context, that is, Item Response Theory (IRT) or Classical Test Theory. Among the latest, one of the most know and widespread method is that based on the Mantel-Haenszel test (Mantel and Haenszel, 1959; Holland and Thayer, 1988), which aims at testing whether there is an association between group membership and item response, conditionally on the total test score. Most specifically, it is based on the analysis of the contingency tables of correct/incorrect (1/0) responses to a given item by two different groups of subjects (e.g., females and males), for the various levels of the total test score. A generalisation of the Mantel-Haenszel approach to multiple groups is due to Penfield (2001), whereas the extension to the case of polytomous items is proposed by Wang and Su (2004).

A further approach for DIF detection is that provided by the estimation of logistic regression models (Swaminathan and Rogers, 1990). Differently from the Mantel-Haenszel approach, logistic regression treats the total test score as a continuous variable and estimates the probability of answering 1 (or 0) to the tested binary item, by using by using the test score, the group membership, and the interaction between these two variables as covariates (Clauser and Mazor, 1998).

In the context of IRT, developed methods to detect item with DIF use ability estimates, in place of the test total score (employed in the Classical Test Theory) and conceptualise DIF in terms of differences in the item parameters estimated separately for each group, commonly named as reference group and focal group. In fact, in the framework of IRT models, item parameters are assumed to be invariant to group membership, so that any difference in the item parameters, estimated separately for each group, indicate the presence of a differential functioning for that item (Bartolucci *et al.*, 2015). Besides, as IRT models allow for a different number of item parameters to be estimated from the data, they allow for the evaluation of DIF for different item properties. Thus, the Rasch model (Rasch, 1961) investigates DIF in the difficulty parameters: differential functioning for a given item is observed if the item characteristic curves, estimated separately for two or more groups, are shifted, so that for one group the conditional probability of endorsing the item is systematically higher (or lower) than that for another group, for all latent trait levels. In this case, the DIF effect is said to be uniform, because the differences between groups in the conditional probabilities are independent of the common latent trait value. Differently from the Rasch model, the 2PL model (two-parameter logistic; Birnbaum, 1968) allows us to detect DIF by looking at the discriminant parameters. This type of DIF corresponds to



ISBN 9788835131557

an interaction between the latent variable and the group membership: DIF is detected anytime the item characteristic curves, estimated separately for two groups (e.g., females and males), have different slopes and cross each other. This implies that the conditional probability to respond correctly to a given item for one group is not constant across the latent trait levels (i.e., it is higher than that for another group for a certain interval of the latent trait values and it is smaller for the remaining values). In this case, the DIF effect is said to be non-uniform. It is worth noticing that the distinction between uniform and non-uniform DIF is not immediately generalisable to polytomous items, because of the presence of a number of non-monotone characteristic curves.

A common drawback for the appropriate use of many DIF detection procedures is the multidimensionality of the data (Gnaldi and Bacci, 2016). In fact, on one side, the choice of total test score as a matching variable (within the Classical Test Theory) is based on the assumption that this score is the most reliable measure of ability. However, if the test does not imply only one dimension, such a score may not be an appropriate choice for comparing groups of examinees. In a similar way, this drawback extends to IRT methods, for which the data must meet the stringent unidimensionality assumption.

Keeping in mind this last key issue, in this work we present a comparison between models for DIF identification, by using the INVALSI data collected through the administration in 2019 of the Mathematics test to pupils of Italian primary schools. In particular, we show an analysis concerning the degree of agreement of the various DIF detection methods developed in the IRT and non-IRT framework, and between unidimensional and multidimensional models. The analysis allows us to assess if and for which items different DIF detection approaches agree/disagree in identifying them as affected by DIF.

The article is organised as follows: in Section 2 we describe some of the main methods employed in the literature to identify DIF, both in the IRT and in the Classical Test Theory frameworks. In Section 3, we report the main results of the analysis of the INVALSI data and the study of the level of agreement between DIF detection models. The main conclusions are drawn in Section 4.

## 2. Material and methods

In this section, the statistical methods for DIF detection are briefly described (Section 2.1), together whit the data used in the application (Section 2.2).



ISBN 9788835131557

## *2.1. DIF detection approaches*

As outlined in the previous section, DIF is an item bias in the measurement of the latent trait. In fact, for an item affected by DIF, the response probability differs between individuals belonging to different groups (e.g., according to gender or geographic location) but with the same ability level. This is the crucial point: subjects are firstly matched according to their ability, then they share the same latent trait level, but the membership to a particular category of group variable causes a shift in response function $P(\,)$.

Formally, we can say that DIF occurs if

$$P(Y|\theta, G = R) \neq P(Y|\theta, G = F),$$

where group variable $G$ assumes two values, generally labelled as reference group ($G = R$) and focal group ($G = F$), $Y$ is the response to a generic item and $\theta$ is the ability (common to both groups).

DIF detection approaches differ as regards the matching variable, since, as already disclosed in the Introduction, it is necessary to compare item response functions for subjects with the same ability level but belonging to different groups. Usually, the total test score (Classical Test Theory) or latent trait estimate (through IRT models) are employed as proxy of ability. As a consequence, methods for DIF detection can be distinguished on the basis of whether they use either Classical Test Theory approach or the IRT approach.

Among non-IRT methods, the one based on Mantel-Haenszel (MH) test, as already said, allows us to verify the presence of association between item response and group membership. It is based on the analysis of contingency tables, obtained by crossing, for each item, the response outcome (correct/incorrect) with group membership of individuals, matched by total test score.

Another non-IRT approach is based on logistic regression: several models are built, considering the following covariates: the total test score, group membership and, eventually, an interaction between them. If group variable has a significant effect, then we are in presence of a uniform DIF, while a non-uniform DIF is detected by a significant interaction between total test score and group variable. These significances may be tested using the usual tests employed in logistic regression, such as Wald test or likelihood ratio test.

Within IRT approaches, DIF detection methods consider the ability estimate in place of the total test score. As known, IRT models are used for analysing latent phenomena starting from an observable manifestation of them. For example, if one is interested in studying mathematical ability, as it is not directly observable by nature, we may exploit responses to a



ISBN 9788835131557

Mathematics assessment test, which represent the observable manifestation of that ability.

IRT models consider the item response probability in function of the respondent's characteristics, generally said latent trait or latent ability, but, also, in function of certain item features, such as its difficulty and discrimination. Among the most widely-used IRT models, we may include the Rasch model (Rasch, 1961), based on the conditional probability of correct response parametrised as follows:

$$logit\ P(Y_{ij} = 1|\theta_i) = \theta_i - \beta_j, \quad (1)$$

where $\beta_j$ is the difficulty parameter and $\theta_i$ is the ability level of student $i$. The Rasch model assumes equal discrimination among items. By removing this constraint, we can use the two-parameter logistic model (2PL; Birnbaum, 1968):

$$logit\ P(Y_{ij} = 1|\theta_i) = \gamma_j(\theta_i - \beta_j),$$

where $\gamma_j$ is the discrimination parameter of item $j$.

A first IRT-based method for DIF detection consists in a simple likelihood ratio (LR) test, which allows us to compare two nested models (Thissen *et al.*, 1988). In this case, the null hypothesis to test is the equality of the item parameters, estimated in both groups (reference and focal). Then, a first IRT model (constrained model) is fitted with identical item parameters for both groups. Afterwards, an augmented model is estimated, in which the parameters of the item suspected to exhibit DIF can be different in the two groups.

LR statistics is obtained as follows:

$$LR = -(l_0 - l_1),$$

where $l_0$ and $l_1$ are the maximised log-likelihood for the constrained and augmented model, respectively. Under the null hypothesis, this test statistic has an asymptotic Chi-square distribution with degrees of freedom equal to the number of unconstrained items in the augmented model with respect to the constrained one.

A second IRT approach is the so-called Lord's method (Lord, 1980), allowing us to test the null hypothesis of no DIF through the direct comparison between the item parameters estimated in the two groups. Such a comparison is based on a test statistic with a Chi-square distribution with degrees of free-



ISBN 9788835131557

dom equal to the number of item parameters included in the model (one for the Rasch model, two for the 2PL model, and so on).

These two IRT approaches refer to classic IRT framework, which assumes unidimensionality. This means that the latent ability underlying the response process is supposed to be unique. However, this assumption cannot be met in real situations, because, when responding to a particular test item, a student activates several sub-abilities, potentially correlated each other and all attributable to a common and more general construct. For example, during the response process to items of a Mathematics test, several mathematical sub-competences are activated, related to mathematical contents and/or cognitive processes (Bartolini Bussi *et al.*, 1999; Douek, 2006; Gnaldi, 2017; Gnaldi and Del Sarto, 2018; Del Sarto, 2019).

Obviously, the ability (latent trait) dimensionality issue becomes critical when aiming at detecting DIF, as it is the matching variable used for testing the presence of this phenomenon. As a consequence, if we consider the ability as unique variable (rather than a multiple one), a bias in subjects' matching would occur, then a potential alteration of results.

### *2.2. Data*

In this work, we use data coming from the INVALSI Mathematics test, administrated in 2019 to pupils of primary school (grade 5). Specifically, only the national sample classes are considered (one for school), in which the test is administrated in presence of an external supervisor. The dataset is therefore related to 24,781 students, belonging to 1,381 classes.

The test at issue is made up of 39 multiple-choice questions, of which the outcome for each student is known, in terms of correct/incorrect response. Another information about test items is their classification, according to mathematical contents and to cognitive processes activated by the student when he/she responds to the item. A further classification criterion has been recently proposed, based on macro-dimensions related to the goals of the National Indications of the first cycle of instruction. In particular, according to this last classification, each question is connected with a goal and goals are grouped further in three macro-dimensions: "Understanding", "Problem solving" and "Reasoning" (INVALSI, 2018).

In this work, we consider this last item classification: in Table 1, together with the observed correct response rate, we show the classification of the 39 items of the INVALSI Mathematics test at issue, according to the National Indication macro-dimensions.



ISBN 9788835131557

*Tab. 1 – Classification of the INVALSI Mathematics test items (grade 5) according to the National Indication macro-dimensions and observed correct response rate (%)*

| *Item* | *Macro-dim* | *%* | *Item* | *Macro-dim* | *%* |
|---|---|---|---|---|---|
| D1 | UND | 77.7 | D16 | UND | 45.3 |
| D2 | PS | 50.4 | D17 | UND | 59.3 |
| D3_a | UND | 81.1 | D18 | PS | 79.3 |
| D3_b | UND | 61.9 | D19 | REAS | 60.6 |
| D4 | UND | 40.3 | D20 | PS | 42.3 |
| D5 | REAS | 60.2 | D21 | UND | 53.6 |
| D6 | UND | 62.0 | D22 | REAS | 56.8 |
| D7 | UND | 54.3 | D23 | PS | 62.5 |
| D8_a | PS | 80.2 | D24 | UND | 48.8 |
| D8_b | PS | 60.0 | D25 | UND | 51.6 |
| D8_c | PS | 38.7 | D26 | UND | 66.9 |
| D9 | UND | 81.2 | D27 | PS | 47.3 |
| D10 | PS | 51.4 | D28 | PS | 51.9 |
| D11 | UND | 32.1 | D29 | REAS | 53.7 |
| D12_a | UND | 75.5 | D30 | UND | 61.7 |
| D12_b | UND | 61.3 | D31 | UND | 42.9 |
| D12_c | UND | 90.2 | D32 | UND | 28.4 |
| D13 | UND | 43.4 | D33 | PS | 76.8 |
| D14 | UND | 35.2 | D34 | UND | 76.2 |
| D15 | PS | 70.1 | | | |

UND: understanding; PS: problem solving; REAS: reasoning.

## 3. Results

This section is devoted to the results about the agreement between approaches for DIF detection. Each method, illustrated in the previous section, identifies a subset of items with DIF and our purpose is to evaluate if and how these approaches agree in the identification of "biased" items. Only with illustrative purposes, in this application we consider gender as group variable. In particular, in the following we show results as regards the agreement in identifying items affected (and not affected) by DIF between the following groups of approaches:

– classical (non-IRT);
– IRT;
– classical *vs* IRT;
– unidimensional *vs* multidimensional IRT.



ISBN 9788835131557

*Tab. 2 – Classical approaches for DIF detection. For each test item, the test statistic is reported, obtained on the basis of a specific approach (Mantel-Haenszel or logistic regression). Moreover, p-values and outcomes (DIF/NO DIF) are also shown*

| | *Mantel-Haenszel* | | | *Logistic* | | |
|---|---|---|---|---|---|---|
| *Item* | *Stat.* | *p-value* | *Outcome* | *Stat.* | *p-value* | *Outcome* |
| D1 | 0.28 | 0.5990 | NO DIF | 0.80 | 0.3724 | NO DIF |
| D2 | 4.19 | 0.0407 | NO DIF | 5.15 | 0.0232 | NO DIF |
| D3_a | 8.07 | 0.0045 | DIF | 9.28 | 0.0023 | DIF |
| D3_b | 11.82 | 0.0006 | DIF | 11.98 | 0.0005 | DIF |
| D4 | 131.69 | < 0.0001 | DIF | 145.87 | < 0.0001 | DIF |
| D5 | 55.78 | < 0.0001 | DIF | 51.31 | < 0.0001 | DIF |
| D6 | 25.23 | < 0.0001 | DIF | 22.17 | < 0.0001 | DIF |
| D7 | 0.11 | 0.7358 | NO DIF | 0.49 | 0.4836 | NO DIF |
| D8_a | 28.34 | < 0.0001 | DIF | 30.00 | < 0.0001 | DIF |
| D8_b | 27.97 | < 0.0001 | DIF | 27.12 | < 0.0001 | DIF |
| D8_c | 17.29 | < 0.0001 | DIF | 13.31 | 0.0003 | DIF |
| D9 | 28.05 | < 0.0001 | DIF | 25.99 | < 0.0001 | DIF |
| D10 | 0.56 | 0.4539 | NO DIF | 1.72 | 0.1892 | NO DIF |
| D11 | 112.85 | < 0.0001 | DIF | 120.02 | < 0.0001 | DIF |
| D12_a | 2.34 | 0.1260 | NO DIF | 2.77 | 0.0960 | NO DIF |
| D12_b | 33.96 | < 0.0001 | DIF | 36.18 | < 0.0001 | DIF |
| D12_c | 19.59 | < 0.0001 | DIF | 24.09 | < 0.0001 | DIF |
| D13 | 136.42 | < 0.0001 | DIF | 154.46 | < 0.0001 | DIF |
| D14 | 0.03 | 0.8735 | NO DIF | 0.27 | 0.6033 | NO DIF |
| D15 | 144.58 | < 0.0001 | DIF | 137.88 | < 0.0001 | DIF |
| D16 | 55.23 | < 0.0001 | DIF | 51.88 | < 0.0001 | DIF |
| D17 | 12.87 | 0.0003 | DIF | 11.81 | 0.0006 | DIF |
| D18 | 63.57 | < 0.0001 | DIF | 68.80 | < 0.0001 | DIF |
| D19 | 94.68 | < 0.0001 | DIF | 98.35 | < 0.0001 | DIF |
| D20 | 104.67 | < 0.0001 | DIF | 111.75 | < 0.0001 | DIF |
| D21 | 2.75 | 0.0973 | NO DIF | 1.45 | 0.2279 | NO DIF |
| D22 | 48.96 | < 0.0001 | DIF | 46.85 | < 0.0001 | DIF |
| D23 | 63.32 | < 0.0001 | DIF | 61.71 | < 0.0001 | DIF |
| D24 | 30.52 | < 0.0001 | DIF | 24.62 | < 0.0001 | DIF |
| D25 | 49.36 | < 0.0001 | DIF | 48.80 | < 0.0001 | DIF |

As far as the first comparison is concerned (between classical methods), approaches based on MH test and on logistic regression are applied to data described in Section 2.2. Then, for each item and for each method, a test statistic is available, which allows us to assess whether that item significantly



ISBN 9788835131557

exhibits DIF (see Table 2). As regards test significance, we use a *p*-value of 1%, instead of the usual 5%, in order to obtain robust results, given the large sample size.

From Table 2 it is possible to build a two-way table about the agreement between the two methods (MH vs. logistic regression). In this new table we report the count of "DIF-positive" cases for both methods, those resulted "negative" according to both, and the count of disagreeing cases, that is, "positive" for one method and "negative" for the other, and vice versa.

As regards the comparison between non-IRT methods, we can refer to Table 3a. As we can see, the two methods perfectly agree, as they both identify 30 items with DIF (out of the 39 test items) and the cells about disagreeing cases are both equal to 0 (i.e., no disagreeing cases).

*Tab. 3 – Agreement between DIF detection approaches: a) non-IRT methods – logistic regression vs. Mantel-Haenszel (MH); b) IRT methods – Lord's test vs. IRT likelihood ratio (IRT-LR) test*

*a)*

| | *MH* | | |
|---|---|---|---|
| *Logistic* | *DIF* | *No DIF* | *Total* |
| DIF | 30 | 0 | 30 |
| No DIF | 0 | 9 | 9 |
| Total | 30 | 9 | 39 |

*b)*

| | *IRT-LR test* | | |
|---|---|---|---|
| *Lord* | *DIF* | *No DIF* | *Total* |
| DIF | 28 | 0 | 28 |
| No DIF | 1 | 10 | 11 |
| Total | 29 | 10 | 39 |

According to the second comparison, related to IRT approaches, in this work we consider only the Rasch model, hence only uniform DIF is investigated. To this aim, we directly show the two-way table about agreement between IRT-based DIF detection methods, that is, Lord's test and IRT-LR test (specific results using Table 2 style are omitted but available upon request). By looking at Table 3b we can note a non-perfect agreement (as instead observed in the previous comparison), however we can assert that the two IRT approaches globally agree in identifying items with DIF (tests agreeing in 38 items out of 39). In fact, only one disagreement is observed,



ISBN 9788835131557

related to item D8_c: IRT-LR test identifies it as affected by DIF, differently from Lord's test.

So far, we have seen an overall agreement within the two lines of DIF detection approaches (non-IRT and IRT), that is, within methods of each approach. It comes natural to wonder what is the agreement between non-IRT and IRT methods. In Table 4 we report the two-way tables about the comparisons between MH method (perfectly agreeing with logistic regression method) and those based on IRT models, Lord's test (Table 4a) and IRT-LR test (Table 4b).

*Tab. 4 – Agreement between non-IRT and IRT DIF detection approaches: a) MH vs. Lord; b) MH vs. IRT likelihood ratio (IRT-LR) test*

*a)*

| | *Lord* | | |
|---|---|---|---|
| *MH* | *DIF* | *No DIF* | *Total* |
| DIF | 28 | 2 | 30 |
| No DIF | 0 | 9 | 9 |
| Total | 28 | 11 | 39 |

*b)*

| | *IRT-LR test* | | |
|---|---|---|---|
| *MH* | *DIF* | *No DIF* | *Total* |
| DIF | 29 | 1 | 30 |
| No DIF | 0 | 9 | 9 |
| Total | 29 | 10 | 39 |

Here too, we can observe an almost-perfect agreement between the two groups of approaches: only two items disagree when comparing MH vs. Lord's test (items D8_c and D17), while results disagree as regards only one item (D17) in the MH vs. IRT-LR test comparison.

Results shown so far are based on unidimensional models, assuming that latent ability underlying item response process is unique and measured by the total test score (non-IRT methods) and by the estimate of the (unique) latent trait (IRT methods). However, as previously mentioned, unidimensionality hypothesis is difficult to meet when the interest is in measuring students' competences.

To this aim, as empirical evidence of this last consideration, a comparison between model fitting is performed, as regards an IRT unidimensional model with respect to its multidimensional counterpart. Specifically, a unidimen-



ISBN 9788835131557

sional Rasch model is fitted at first (eq. 1); then, a multidimensional Rasch model is considered, whose dimensionality structure is specified according to the National indications macro-dimensions (Table 1), then based on the three dimensions "Understanding", "Problem solving" and "Reasoning".

Results about this comparison (fitting of unidimensional and multidimensional Rasch models) are reported in Table 5, from which we can deduce a better fitting of the multidimensional model, in terms of both the information criteria – Bayesian Information Criterion (BIC; Schwartz, 1978) and Akaike Information Criterion (AIC; Akaike, 1973) – and the likelihood ratio test ($p$-value < 0.0001).

*Tab. 5 – Comparison between unidimensional Rasch model and its multidimensional counterpart (based on the three National indications macro-dimensions, i.e., "Understanding", "Problem solving" and "Reasoning"): maximised log-likelihood (log-lik), number of model parameters (#par), information criteria (AIC and BIC), LR test statistic, degrees of freedom and p-value*

| | *log-lik* | *#par* | *AIC* | *BIC* | *LR test statistic* | *Degrees of freedom* | *p-value* |
|---|---|---|---|---|---|---|---|
| Unidim. | -549,130.9 | 40 | 1,098,342 | 1,098,667 | | | |
| Multidim. | -549,011.7 | 48 | 1,098,119 | 1,098,509 | 238.4 | 8 | < 0.001 |

Due to the last evidence, it is reasonable to assume a three-dimension structure for the data at issue, in particular as regards the ability underlying the test considered here (i.e., the mathematical ability of grade 5 students). We can therefore proceed with the analysis of the agreement between IRT-based DIF detection methods, distinguished by the dimensionality assumption: unidimensional *vs* multidimensional approach, both based on IRT-LR test. To this aim, we can look at Table 6, reporting the agreement between these two approaches.

*Tab. 6 – Agreement in DIF detection, related to the comparison between unidimensional and multidimensional IRT approaches*

| | *Unidim. IRT* | | |
|---|---|---|---|
| *Multidim. IRT* | *DIF* | *No DIF* | *Total* |
| DIF | 23 | 5 | 28 |
| No DIF | 6 | 5 | 11 |
| Total | 29 | 10 | 39 |

By looking at the marginal totals, we can note that, according to a unidimensional model, it is possible to identify 29 items (out of 39) with DIF,



ISBN 9788835131557

while its multidimensional counterpart detects one less (28). However, by inspecting the joint frequencies about agreement/disagreement, we can observe that results do no agree as regards 11 items, equal to around 28% of the total test items. In fact, six items are detected as "DIF-positive" according to the unidimensional IRT approach but "DIF-negative" on the basis of the multidimensional version, while the opposite holds for five items.

By looking over the items for which we have disagreement, we can notice that, among the six items that are "positive" to unidimensional test but "negative" to the multidimensional one (D3_a, D5, D8_b, D19, D22, D29), all the four items belonging to the "Reasoning" dimension are present. Furthermore, among the five items for which the opposite holds, that is, "positive" to multidimensional test and "negative" to the unidimensional one (D10, D12_a, D17, D21, D30), almost all of them belong to the "Understanding" dimension, except item D10.

## 4. Conclusions

The present contribution aims at comparing different approaches for Differential Item Functioning (known as DIF) detection, which occurs anytime the response probability to a given item differs between groups of examinees with the same ability level (e.g., groups defined according to gender, geographic location, etc.).

In particular, we compare different DIF detection methods, developed both in the Classical Test Theory framework and in the Item Response Theory framework, and based on alternative hypothesis: that of unidimensionality and that of multidimensionality of the latent trait underlying the response process.

The compared methods have been applied to data collected through the administration in 2019 of the INVALSI Mathematics test to pupils of grade 5, belonging to Italian primary schools. As known, the items of the INVALSI test administered in 2019 can be classified differently according to different criteria, such as on the basis of their mathematical content, the cognitive processes they activate, and their goals as specified in the National Indications of the first cycle of instruction: "Understanding", "Problem Solving" and "Reasoning".

The results show that, when identifying items affected by DIF, the various DIF detection methods employed in this study tend to agree both when considering a comparison between methods belonging to the same framework (Classical Test Theory or Item Response Theory) and a comparison



ISBN 9788835131557

of approaches across frameworks (Classical Test Theory vs. Item Response Theory).

However, the introduction in the model of the multidimensional structure of the test at issue – which we recover by relying on the National Indications – leads us to observe an important disagreement in the results when comparing the unidimensional IRT model for DIF detection and its multidimensional counterpart. Such a disagreement concerns 11 items overall, that is, around 30% of the test items. Besides, we observe that the four test items classified as "Reasoning" by the National indications are all included in the group of items for which DIF results based on unidimensional ad multidimensional IRT models do not agree. Specifically, the items at issue are items D5, D19, D22 and D29. In this regard, it is worth being noticed that great attention has been given to this specific sub-competence of the Mathematics ability. In fact, in respect to the development of transversal competences, they "are relevant for the formation of an active and conscious citizenship, in which each person is ready to listen attentively and critically to the other and to a comparison based on relevant and pertinent topics. In particular, education in argumentation (reasoning) can be an antidote to the proliferation of false or uncontrollable information" (INVALSI, 2018, p. 8).

Consequently, using an approach that does not take into account the multidimensionality of the test leads to an erroneous consideration of these four items. Specifically, in the applicative example provided in this paper, using a one-dimensional approach, we would be led to conclude that the four items at issue are affected by DIF (which results, in particular, in significantly higher difficulty parameters for females than for males) and which could therefore affect the overall validity of the test, with regard to the specific sub-competence "Reasoning". On the other hand, when the multidimensional structure of the test at hand is taken into account, the four items are no longer identified as suffering from DIF. A situation like the one just described, in phase of validation of a test, can therefore lead to not reliable considerations, that is to say, to claim the bias of "false positive" items, and, on the other hand, to exclude the bias with respect to "false negative" items, with consequent corrective actions directed to good items, while neglecting items showing criticality.

It is therefore essential that, in the phase of validiation of a test and its individual items, different approaches for DIF detection – based on different methodological hypotheses and taken from both the Classical Test Theory context and the IRT framework – are concurrently employed. This is important in order to verify their convergence towards similar results and to get information as to which items are systematically affected by DIF under all methods or, oth-



ISBN 9788835131557

erwise, only under some of them. This first indication will be the ground for the first evaluations by experts on the relative goodness of the items.

However, this preliminary phase should be accompanied by a dimensionality check of the test because, as stressed several times in this work and in other previous papers, INVALSI Mathematics tests do not imply a one-dimensional latent variable but a multidimensional ability. Ignoring its multidimensionality leads to erroneous considerations also in the process of detection of items affected by DIF. If, like in the case of the study presented here, there is evidence of better fit to the data of multidimensional IRT models compared to unidimensional counterparts, it is essential that DIF detection methods also accounting for multidimensionality are used in place of (or next to) one-dimensional DIF detection methods. This last condition is necessary to have a picture as complete as possible and to address corrective actions only with respect to those items that are unequivocally classified as biased under all methods or under multidimensional methods alone.

Finally, this work may be developed by considering an alternative approach based on the triplet DBI (acronym for DIF, Bias and Impact), which aims at disentangling DIF from other strictly-related concepts, such as bias and impact (in this regard, see, for example, Wu *et al.*, 2017).

## References


Akaike H. (1973), "Information theory and an extension of the maximum likelihood principle", in B.N. Petrov, F. Csáki (*eds.*), *Proceedings of the second international symposium of information theory*, Akadémiai Kiado, Budapest.

Bartolini Bussi M.G., Boni M., Ferri F., Garuti R. (1999), "Early approach to theoretical thinking: gears in primary school", *Educational Studies in Mathematics*, 39, 1, pp. 67-87.

Bartolucci F., Bacci S., Gnaldi M. (2015), *Statistical analysis of questionnaires: A unified approach based on R and Stata*, CRC Press, Boca Raton.

Birnbaum A. (1968), "Some latent trait models and their use in inferring an examinee's ability", in F.M. Lord, M.R. Novick (*eds.*), *Statistical theories of mental test scores*, Addison-Wesley, Reading.

Clauser B.E., Mazor K.M. (1998), "Using statistical procedures to identify differentially functioning test items", *Educational Measurement: Issues and Practic*e, 17, 1, pp. 31-44.

Del Sarto S. (2019), "Un'analisi sulla bontà di adattamento di tre modelli IRT multidimensionali ai dati INVALSI di Matematica", in P. Falzetti (a cura di), *Implementazione e miglioramento del dato. II Seminario "I dati INVALSI: uno strumento per la ricerca"*, FrancoAngeli, Milano, pp. 34-50.




ISBN 9788835131557

Douek N. (2006), "Some remarks about argumentation and proof", in P. Boero (*ed.*), *Theorems in school: from history, epistemology and cognition to classroom practice*, Sense Publishers, Rotterdam.
Gnaldi M. (2017), "A multidimensional IRT approach for dimensionality assessment of standardised students' tests in mathematics", *Quality & Quantity*, 51, 3, pp. 1167-1182.
Gnaldi M., Bacci S. (2016), "Joint assessment of the latent trait dimensionality and observed differential item functioning of students' national tests", *Quality & Quantity*, 50, 4, pp. 1429-1447.
Gnaldi M., Del Sarto S. (2018), "Variable weighting via multidimensional IRT models in Composite Indicators construction", *Social Indicators Research*, 136, 2, pp. 1139-1156.
Holland P.W., Thayer D.T. (1988), "Differential item performance and the Mantel-Haenszel procedure", in H. Wainer, H.I. Braun (*eds.*), *Test Validity*, Erlbaum, Hillsdale.
INVALSI (2018), *Quadro di Riferimento delle prove INVALSI di Matematica*, retrieved on April 6, 2021, from: https://INVALSI-areaprove.cineca.it/docs/file/QdR_MATEMATICA.pdf.
Lord F.M. (1980), *Applications of Item Response Theory to practical testing problems*, Erlbaum, Hillsdale.
Mantel N., Haenszel W. (1959), "Statistical aspects of the analysis of data from retrospective studies of disease", *Journal of the National Cancer Institute*, 22, 4, pp. 719-748.
Penfield R.D. (2001), "Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures", *Applied Measurement in Education*, 14, 3, pp. 235-259.
Rasch G. (1961), "On general laws and the meaning of measurement in psychology", in J. Neyman (*ed.*), *Proceedings of the IV Berkeley symposium on mathematical statistics and probability*, University of California Press, Berkeley.
Schwarz G. (1978), "Estimating the dimension of a model", The Annals of Statistics, 6, 2, pp. 461-464.
Swaminathan H., Rogers H. (1990), "Detecting differential item functioning using logistic regression procedures", *Journal of Educational Measurement*, 27, 4, pp. 361-370.
Thissen D., Steinberg L., Wainer H. (1988), "Use of item response theory in the study of group differences in trace lines", in H. Wainer , H.I. Braun (*eds.*), *Test Validity*, Erlbaum, Hillsdale.
Wang W.C., Su Y.Y. (2004), "Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items", *Applied Psychological Measurement*, 28, 6, pp. 450-480.
Wu A.D., Liu Y., Stone J.E., Zou D., Zumbo B.D. (2017), "Is difference in measurement outcome between groups differential responding, bias or disparity? A methodology for detecting bias and impact from an attributional stance", *Frontiers in Education*, 2, p. 39.



ISBN 9788835131557

# *5. Cross-cohort changes in indicators of tolerance among Italian youth*

by Maria Magdalena Isac, Laura Palmerio, Elisa Caponera

Tolerance, generally defined as positive feelings toward diversity as well as an understanding and endorsement of equality between different groups (Cote and Erikson, 2009), is considered an important democratic attitude and an essential prerequisite for a peaceful coexistence in the increasingly diverse contemporary societies (Freitag and Rapp, 2015). In the Italian context challenged by unprecedented migration, monitoring and promoting tolerance in schools is an essential part of policies focused on inclusive citizenship education and intercultural dialogue. Therefore, comparative studies focused on identifying patters of change in young people's tolerant attitudes are highly needed. In this research, we argue that comparability must be empirically assessed and ensured for the measurement of relevant indicators that serve to monitor cross-cohort changes in indicators of tolerance among italian youth. To this end, we aim to a) evaluate the extent to which the scales of tolerance toward equal rights for immigrants, ethnic groups, and women are measurement invariant in two cohorts (2009; 2016) and b) explore how patterns of change in tolerant attitudes vary by cohort. Using the framework and data provided by the International Civic and Citizenship Education Studies (ICCS, 2009; 2016) (Schulz *et al.*, 2010; 2018; Torney-Purta *et al.*, 2001) conducted by the International Association for the Evaluation of Educational Achievement (IEA) and coordinated for Italy by the International Large-Scale Assessments Unit of the Italian Institute for the Evaluation of the Education System (INVALSI), we examine the extent to which average comparisons of cross-cohort differences in young people's tolerant attitudes toward immigrants, ethnic/racial groups, and gender equality are empirically justified. Multiple-group confirmatory factor analysis (MGCFA) (Joreskog, 1971) is applied to estimate the three-dimensional measurement model of the concept and test its measurement invariance across the two cohorts. Results



ISBN 9788835131557

of MGCFA pointed out that cross-cohort comparability cannot be achieved at the highest level of measurement invariance, i.e. scalar invariance, but reaches the configural and metric levels of invariance. Assuming the proposed measurement model, the factor scores (scale means) on these variables cannot be compared with confidence across the two cohorts. However, we find that at both measurement points students tend to give similar meaning to these concepts and tend to respond to the items in the same way (metric invariance). The implications of these findings are that cross-cohort comparisons are to be interpreted with caution. Nevertheless, the data from both measurement points can be useful in analyses conducted across both datasets that may seek to explore, for example, associations between these concepts and other theoretical constructs of interest.

*La tolleranza, generalmente definita come un sentimento positivo nei confronti della diversità e come comprensione e sostegno all'uguaglianza tra i diversi gruppi (Cote ed Erikson, 2009), è considerata un importante atteggiamento democratico e un prerequisito essenziale per una coesistenza pacifica nelle società contemporanee sempre più diverse (Freitag e Rapp, 2015). Negli ultimi anni in Italia si sta assistendo a un fenomeno di migrazione senza precedenti. Il monitoraggio e la promozione della tolleranza nelle scuole è dunque una parte essenziale delle politiche incentrate sull'educazione alla cittadinanza inclusiva e sul dialogo interculturale. Per questo motivo, sono particolarmente necessari studi comparativi volti a individuare i modelli di cambiamento degli atteggiamenti tolleranti dei giovani. In questa ricerca, sosteniamo che la comparabilità deve essere valutata empiricamente in modo da garantire la misurazione di fattori rilevanti che servono a monitorare, tra coorti di studenti, i cambiamenti degli indicatori di tolleranza tra i giovani italiani. A tal fine, ci proponiamo di: a) valutare in che misura le scale di tolleranza verso la parità di diritti per gli immigrati, i gruppi etnici e le donne sono invarianti in due coorti (2009; 2016); b) esplorare come i modelli di cambiamento degli atteggiamenti tolleranti variano da coorte a coorte. Nel presente studio, utilizzando i framework e i dati forniti dagli International Civic and Citizenship Education Studies (ICCS, 2009; 2016) (Schulz et al., 2010; 2018; Torney-Purta et al., 2001) condotti dall'International Association for the Evaluation of Educational Achievement (IEA) e coordinati per l'Italia dall'Area Indagini internazionali dell'INVALSI, si esamina in che misura i confronti medi delle differenze cross-coorte negli atteggiamenti tolleranti verso gli immigrati, i gruppi etnici/razziali e l'uguaglianza di genere dei giovani siano giustificati empiricamente. È stata condotta un'analisi fattoriale confermativa multigruppo (MGCFA) (Joreskog, 1971) per stimare il*



ISBN 9788835131557

*modello di misura a tre dimensioni del costrutto e testarne l'invarianza fra le due coorti. I risultati della MGCFA hanno evidenziato che la comparabilità tra coorti non può essere ottenuta al più alto livello di invarianza di misura, cioè l'invarianza scalare, ma raggiunge i livelli di invarianza configurale e metrica. Assumendo il modello di misurazione proposto, il confronto tra le due coorti dei punteggi fattoriali (medie di scala) su queste variabili non è affidabile. Tuttavia, in entrambi i punti di misurazione gli studenti tendono a dare un significato simile a questi concetti e tendono a rispondere agli item nello stesso modo (invarianza metrica). Le implicazioni di questi risultati sono che i confronti tra coorti devono essere interpretati con cautela. Tuttavia, i dati di entrambi i punti di misurazione possono essere utili nelle analisi comparate su entrambi i set di dati per esplorare, per esempio, le associazioni tra questi concetti e altri costrutti teorici di interesse.*

## 1. Background

Tolerance, generally defined as positive feelings toward diversity as well as an understanding and endorsement of equality between different groups (Cote and Erickson, 2009), is considered an important democratic attitude and an essential prerequisite for a peaceful coexistence in the increasingly diverse contemporary societies (Freitag and Rapp, 2015).

In the Italian context challenged by unpreceded migration, monitoring and promoting tolerance in schools is an essential part of policies focused on inclusive citizenship education and intercultural dialogue. Therefore, comparative studies focused on identifying patters of change in young people's tolerant attitudes are highly needed.

Data for Italy and several other European countries regarding these aspects are available from the International Civic and Citizenship Education Studies (ICCS, 2009; 2016) conducted by the International Association for the Evaluation of Educational Achievement (IEA) and coordinated for Italy by the International Large-Scale Assessments Unit of the Italian Institute for the Evaluation of the Education System (INVALSI). These studies use questionnaires and inquire into young people's beliefs about equal rights and opportunities for different groups in society based on gender, ethnic/racial status and immigration background and give the possibility to evaluate the change in average scores in these attitudes over time (Schulz *et al.*, 2018; Schulz, Ainley and Fraillon, 2011).

Nevertheless, if latent factor means are to be meaningfully compared across time, the construct needs to be understood and operationalized in a



ISBN 9788835131557

similar way in each measurement point (Davidov et al., 2014; Rutkowski and Svetina, 2017; van de Vijver and Tanzer, 2004). For this reason, secondary users of data collected in such studies are urged to test the assumption of comparability or measurement invariance (French and Finch, 2006; Jöreskog, 1971; Putnick and Bornstein, 2016). To this end, in this research, we aim to evaluate the extent to which the scales of tolerance toward equal rights for immigrants, ethnic groups, and women are measurement invariant in two cohorts of the ICCS study (2009 and 2016) in Italy. We do so elaborating on a measurement model identified in previous research that was found to be comparable among the European countries participating in ICCS 2016 (see Isac, Palmerio and van der Werf, 2019).

## 2. Method

We used data from two cohorts of ICCS, 2009 and 2016 for Italy, were the main data source for all the analyses. In each cohort, the surveyed students are representative samples of the population of grade 8 students. More specifically, the studies followed a two-stage cluster sampling strategy. In a first stage probability proportional to size (PPS) procedures were used to select schools. In the second stage, within each sampled school, an intact class from the target grade was selected at random, with all the students in this class participating in the study. 3,357 Italian students participated in ICCS 2009 and 3,446 students in ICCS 2016.

Based on ICCS 2009 and 2016 data, the construct of tolerance was measured as young people's beliefs about equal political and cultural rights and opportunities for (three) different groups in society based on immigration background, ethnic/racial status and gender (Schulz *et al.,* 2011; 2018). Three scales are used to measure this three-dimensional construct: a) student attitudes toward equal rights for immigrants, b) student attitudes toward equal rights for all ethnic/racial groups, and c) student attitudes toward gender equality. The variables and items used as indicators for the three dimensions of "attitudes toward equal rights" are described in Table 1. Each construct is captured by a set of items measured on 4-point Likert scales ranging from strongly disagree to strongly agree. Some of the items were reverse coded to ensure that high scores on each item reflect positive attitudes toward the three groups.



ISBN 9788835131557

*Tab. 1 – Measures of attitudes toward equal rights for ethnic groups, women and immigrants*

| *Item code* | *Item text* |
|---|---|
| *Domain 1: Attitudes toward equal rights for all ethnic/racial groups* | |
| IS3G25A* | All <ethnic/racial groups> should have an equal chance to get a good education in <country of test>. |
| IS3G25B* | All <ethnic/racial groups> should have an equal chance to get good jobs in <country of test>. |
| IS3G25C* | Schools should teach students to respect <members of all ethnic/racial groups>. |
| IS3G25E* | <Members of all ethnic/racial groups> should have the same rights and responsibilities. |
| *Domain 2: Attitudes toward gender equality* | |
| IS3G24C | Women should stay out of politics. |
| IS3G24D | When there are not many jobs available, men should have more right to a job than women. |
| IS3G24F | Men are better qualified to be political leaders than women. |
| *Domain 3: Attitudes toward equal rights for immigrants* | |
| ES3G04B* | <Immigrant> children should have the same opportunities for education that other children in the country have |
| ES3G04C* | <Immigrants> who live in a country for several years should have the opportunity to vote in elections |
| ES3G04D* | <Immigrants> should have the opportunity to continue their own customs and lifestyle |
| ES3G04E* | <Immigrants> should have the same rights that everyone else in the country has |

* = Item reversed coded.

Data preparation was done with the IEA IDB analyzer (IEA, 2017) and IBM SPSS Statistics 23.00 (IBM Corp., 2015). All measurement invariance analyses were performed in Mplus 7.4 (Muthén and Muthén, 2017) taking into account the complex survey design of the ICCS studies. To handle missing data, we used the full information maximum likelihood (FIML) method implemented in Mplus 7.4.

To establish if average scores on tolerance toward equal rights for immigrants, ethnic groups, and women are comparable across the two ICCS cohorts, measurement invariance was investigated in a factor analytical framework. More specifically, we applied multiple-group confirmatory factor analysis (MGCFA) (Jöreskog, 1971; Steenkamp and Baumgartner, 1998) in which any parameter in the factor analysis models (factor loadings, factor variances, factor covariances, and unique variances to assess whether comparisons of average scale scores across the two measurement points can



ISBN 9788835131557

be made with confidence. In order to address the ordered categorical character of the data (4 point Likert scale), we specified a CFA model that estimates polychoric correlations and asymptotic covariance matrices to reflect the relations between response variables with a weighted least square mean variance (WLSMV) estimator. We tested a first-order correlated three-factor model of attitudes toward equal rights encompassing a) student attitudes toward equal rights for all ethnic/racial groups, b) student attitudes toward gender equality, and, c) student attitudes toward equal rights for immigrants.

The assessment of measurement invariance involved the comparison of the three nested competing models, i.e. configural, metric and scalar models (Brown, 2014; Putnick and Bornstein, 2016). The configural invariance model tested if the instrument measures the same latent factors and if the set of items associated with each factor is similar across measurements. The metric invariance model tested whether the factors have the same meaning and the same measurement unit in both groups. The scalar invariance model tested, in addition to equal item loadings, that item thresholds (the levels of the categorical items) are equal in both groups. Reaching the level of scalar measurement invariance was taken as an indication that valid cross-country comparisons of factor scores (scale means) are defensible. For model fit evaluation, we observed the following guidelines: RMSEA $\leq 0.060$; CFI $\geq 0.950$; TLI $\geq 0.950$; as well as $\Delta$CFI, $\Delta$RMSEA $< 0.01$ for nested models comparisons (Brown, 2014).

## 3. Results

### *3.1. Cohort-specific models*

In a preliminary step we tested by means of confirmatory factor analysis (CFA) the first-order correlated three-factor model of attitudes toward equal rights in each of the ICCS samples, i.e. ICCS 2009 and ICCS 2016.

The results (see Tables 2 and 3) of the cohort-specific CFA models estimated on the ICCS 2009 and 2016 data indicate that the model showed an adequate fit in both samples. The same number of (three) correlated factors with similar patterns of item loadings was identified in both cohorts. Specifically, item loadings, were well above the 0.600 for all scales and in both cohorts, ranging from 0.635 to 0.822 (see Table 3). Fit indices largely fell within acceptable ranges with RMSEA values below 0.060 and CFI and TLI well above 0.950.



ISBN 9788835131557

*Tab. 2 – Results of confirmatory factor analysis. Cohort-specific models*

| | | *First-order correlated three-factor model* | | |
|---|---|---|---|---|
| *ICCS study* | *N* | *RMSEA* | *TLI* | *CFI* |
| ICCS 2009 | 3,357 | 0.047 | 0.980 | 0.985 |
| ICCS 2016 | 3,446 | 0.053 | 0.976 | 0.982 |

Note. N = sample size, RMSEA = Root Mean Square Error of Approximation, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index.
Cut-off criteria for model fit evaluation: RMSEA < 0.06; CFI > 0.95; TLI > 0.95.

*Tab. 3 – Results of confirmatory factor analysis. Item loadings*

| *Standardized item loadings* | *ICCS 2009* | *ICCS 2016* |
|---|---|---|
| *Domain 1: Attitudes toward equal rights for all ethnic/racial groups* | | |
| IS3G25A | 0.849* | 0.860* |
| IS3G25B | 0.852* | 0.872* |
| IS3G25C | 0.774* | 0.761* |
| IS3G25E | 0.820* | 0.839* |
| *Domain 2: Attitudes toward gender equality* | | |
| IS3G24C | 0.749* | 0.804* |
| IS3G24D | 0.794* | 0.809* |
| IS3G24F | 0.837* | 0.818* |
| *Domain 3: Attitudes toward equal rights for immigrants* | | |
| ES3G04B | 0.868* | 0.882* |
| ES3G04C | 0.686* | 0.708* |
| ES3G04D | 0.647* | 0.635* |
| ES3G04E | 0.864* | 0.863* |

* p < .001.

## *3.2. Results of multiple-group analysis*

The results (see Table 4) at the configural and metric levels of invariance largely comply with the model fit evaluation criteria both in terms of overall fit indices (e.g. RMSEA ≤ 0.060; CFI ≥ 0.950; TLI ≥ 0.950) as well as comparative fit (ΔCFI, ΔRMSEA <0.01). Nevertheless, when comparing the fit of the scalar model to the one of the metric model the results do not support the assumption of scalar invariance (ΔCFI, ΔRMSEA > 0.01) indicating that the three factors have the same meaning and the same measurement unit in both groups but item thresholds (the levels of the categorical items) are not equal in both measurement moments.



ISBN 9788835131557

*Tab. 4 – Results of multiple-group analysis, overall model*

| *Model* | *Full sample* | *RMSEA* | *CFI* | *TLI* |
|---|---|---|---|---|
| M1 | Configural | 0.045 | 0.990 | 0.986 |
| M2 | Metric | 0.041 | 0.990 | 0.989 |
| M3 | Scalar | 0.043 | 0.985 | 0.987 |
| | Nested models comparisons | ΔRMSEA | ΔCFI | |
| | Metric vs configural | 0.004 | 0.000 | |
| | Scalar vs metric | -0.002 | 0.005 | |

Note. RMSEA = Root Mean Square Error of Approximation, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, ΔRMSEA = change in RMSEA; ΔCFI = change in CFI.

## 4. Conclusion

In the current research we evaluated the extent to which the scales of tolerance toward equal rights for immigrants, ethnic groups, and women are measurement invariant in two cohorts of the ICCS study (2009 and 2016) in Italy. To this end, we examined the extent to which average comparisons of cross-cohort differences in young people's tolerant attitudes toward immigrants, ethnic minorities and women in the context of the ICCS 2009 and 2016 study are justified.

Results of multiple-group confirmatory factor analysis pointed out that cross-cohort comparability cannot be achieved at the highest level of measurement invariance, i.e. scalar invariance but reaches the configural and metric levels of invariance. Assuming the proposed measurement model, the factor scores (scale means) on these variables cannot be compared with confidence across the two cohorts. However, we find that at both measurement points students tend to give similar meaning to these concepts and tend to respond to the items in the same way (metric invariance). The implications of these findings are that cross-cohort comparisons are to be interpreted with caution. Nevertheless, the data from both measurement points can be useful in analyses conducted across both datasets that may seek to explore, for example, associations between these concepts and other theoretical constructs of interest.

Future research could involve further analyses that could go beyond the assumption of full measurement invariance by redefining the construct (e.g. omitting some of the items and retesting the models) or seek only partial measurement invariance (Byrne and van de Vijver, 2014; Marsh *et al.*, 2017; Putnick and Bornstein, 2016).



ISBN 9788835131557

## References


Brown T.A. (2014), *Confirmatory Factor Analysis for Applied Research. Methodology in the Social Sciences*, Guilford, London.

Byrne B.M., van de Vijver F.J.R. (2014), "Factorial Structure of the Family Values Scale From a Multilevel-Multicultural Perspective", *International Journal of Testing*, 14, 2, pp. 168-192.

Côté R.R., Erickson B.H. (2009), "Untangling the Roots of Tolerance", *American Behavioral Scientist*, 52, 12, pp. 1664-1689, retrieved on April 6, 2021, from: http://journals.sagepub.com/doi/10.1177/0002764209331532.

Davidov E., Meuleman B., Cieciuch J., Schmidt P., Billiet J. (2014), "Measurement Equivalence in Cross-National Research", *Annual Review of Sociology*, 40, 1, pp. 55-75.

Freitag M., Rapp C. (2015), "The Personal Foundations of Political Tolerance towards Immigrants", *Journal of Ethnic and Migration Studies*, 41, 3, pp. 351-373.

French B.F., Finch W.H. (2006), "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance", *Structural Equation Modeling: A Multidisciplinary Journal*, 13, 3, pp. 378-402, retrieved on April 6, 2021, from: http://www.tandfonline.com/doi/abs/10.1207/s15328007sem1303_3.

Isac M.M., Palmerio L., van der Werf M.P.C. (Greetje) (2019), "Indicators of (in) tolerance toward immigrants among European youth: an assessment of measurement invariance in ICCS 2016", *Large-Scale Assessments in Education*, 7, 1, p. 6.

Jöreskog K.G. (1971), "Simultaneous factor analysis in several populations", *Psychometrika*, 36, 4, pp. 409-426.

Marsh H.W., Guo J., Parker P.D., Nagengast B., Asparouhov T., Muthén B., Dicke T. (2017), "What to do When Scalar Invariance Fails: The Extended Alignment Method for Multi-Group Factor Analysis Comparison of Latent Means Across Many Groups", *Psychological Methods*, retrieved on April 6, 2021, from: http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000113.

Putnick D.L., Bornstein M.H. (2016), "Measurement invariance conventions and reporting: The state of the art and future directions for psychological research", *Developmental Review*, 41, pp. 71-90, retrieved on April 6, 2021, from: https://www.sciencedirect.com/science/article/pii/S0273229716300351.

Rutkowski L., Svetina D. (2017), "Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance", *Applied Measurement in Education*, 30, 1, pp. 39-51, retrieved on April 6, 2021, from: http://dx.doi.org/10.1080/08957347.2016.1243540.

Schulz W., Ainley J., Fraillon J. (2011), *ICCS 2009 Technical Report*, International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands, 2010.

Schulz W., Carstens R., Losito B., Fraillon J. (2018), *ICCS 2016 Technical Report*, Amsterdam.




ISBN 9788835131557

Steenkamp J.E.M., Baumgartner H. (1998), "Assessing Measurement Invariance in Cross-National Consumer Research", *Journal of Consumer Research*, 25, 1, pp. 78-107, retrieved on April 6, 2021, from: https://academic.oup.com/jcr/article-lookup/doi/10.1086/209528.
van de Vijver F., Tanzer N.K. (2004), "Bias and equivalence in cross-cultural assessment: An overview", *Revue Europeenne de Psychologie Appliquee*, 54, 2, pp. 119-135.



ISBN 9788835131557

# *6. Automated assessment of open-ended question of INVALSI tests*

by Michele Marsili, Cecilia Bagnarol, Silvia Donno, Emiliano Campodifiori

This work describes the new procedures of automated corrections of free-form answers given by the 8th, 10th and 13th grade students to open-ended questions in CBT (Computer Based Test) INVALSI tests. INVALSI team, composed of statistical and computer scientists, responsible of open-ended question correction, has implemented an algorithm to process text strings of different complexity.

Before survey distribution, the correction team and the items authors group discuss to define the correction criteria, that is a set of rules to determine the correct or incorrect classification for each answer given by the students for a specific item. The discussion produced, moreover, the indications to remove useless elements for the classification, then translated in operations of the algorithm on the textual data such as punctuation detection and removal, special characters, articles, conjunctions, word lemmatisation, etc.

The answer strings were subsequently processed by a "data cleaning" operation, that was focused on the automated correction of spelling and typing errors, by detection and substitution of "out-of-vocabulary" words (OOV words).

After the "data cleaning" phase, the correction criteria fixed by the experts have been translated in logical IT patterns, aiming to uniquely defining the set of admissible ways to give a correct answer.

The last test phases of the algorithm were characterized by a constant exchange of information about the encoding, among the authors' team and the correction team, this passage being critical to refine the logical rules used for correction and to get more consistency and precision between the encoding produced by the algorithm and the authors' indications.

The final test of the algorithm ends with a comparison between the manual encoding by video correction and the one processed by the algorithm on



ISBN 9788835131557

a set of items already processed in a former test: the algorithm is accounted as accurate enough and aligned to the indications of authors' team when the complete accordance of the two encoding was achieved.

The methodological approach, countable as a method of supervised automated correction, represents a valid compromise between a manual encoding and a totally automated one, typical of the machine learning algorithms. This method has indeed the benefit of considerably reduce the hours/man needed to correct the open-ended answer items, when compared to a manual procedure, and get a better accuracy reducing the wrong encoding matches, when compared to a non-supervised automated procedure.

A comparison between supervised and non-supervised automated procedure has been eventually done to evaluate the distance between the two methodological approaches.

*Il presente lavoro di ricerca illustra le nuove procedure di correzione automatizzata delle risposte aperte, introdotte per l'a.s. 2018/19, delle prove INVALSI di Italiano e Matematica somministrate in modalità CBT (Computer Based Test) agli studenti dei gradi 8 (terza secondaria di primo grado), 10 (seconda secondaria di secondo grado) e 13 (quinta secondaria di secondo grado). Il team INVALSI di correzione delle risposte aperte, costituito da statistici ed informatici, ha implementato un algoritmo per il trattamento di stringhe di testo più o meno articolate.*

*Nella fase che precede la somministrazione delle prove, il team di correzione si è confrontato con il gruppo degli autori degli item per definire i criteri di correzione, ovvero una serie di regole che determinano la classificazione in corretta o errata per ciascuna risposta data dagli studenti a un determinato item. Dal confronto sono emerse, inoltre, le indicazioni per la rimozione di elementi non utili alla classificazione tradotte poi in operazioni dell'algoritmo sui dati testuali come l'individuazione e rimozione della punteggiatura, dei caratteri speciali, degli articoli, delle congiunzioni, la lemmatizzazione delle parole ecc.*

*L'operazione successiva di "data cleaning" cui vengono sottoposte le stringhe di risposta è centrata invece sulla correzione automatizzata degli errori di ortografia e digitazione tramite l'individuazione e la sostituzione delle parole "fuori vocabolario" (*OOV words*).*

*Conclusa la fase di "data cleaning", i criteri di correzione stabiliti dal gruppo di esperti sono stati tradotti in pattern logici informatici volti a definire univocamente l'insieme di modi ammissibili di fornire una risposta corretta.*

*Le fasi di collaudo dell'algoritmo sono state caratterizzate da un costante scambio di informazioni sulla codifica fra il team degli autori e il team*



ISBN 9788835131557

*di correzione, passaggio questo cruciale per affinare le regole logiche di correzione utilizzate e di conseguenza per ottenere una sempre maggiore coerenza e precisione fra la codifica prodotta dall'algoritmo e le indicazioni degli autori.*

*Il collaudo dell'algoritmo di correzione si conclude mediante un confronto tra la codifica prodotta manualmente mediante correzione "a video" e quella elaborata dall'algoritmo su un insieme di item oggetto di un precedente pre-test: al verificarsi della perfetta concordanza tra le due codifiche l'algoritmo è ritenuto sufficientemente preciso e allineato alle indicazioni del team di autori.*

*L'approccio metodologico utilizzato, da annoverarsi tra i metodi di correzione automatizzata supervisionata, rappresenta un valido compromesso tra una codifica manuale e una totalmente automatizzata tipica degli algoritmi di machine learning. L'utilizzo di questa metodologia, infatti, ha il vantaggio di ridurre sensibilmente le ore/uomo necessarie allo svolgimento della correzione degli item a risposta aperta, se paragonata a una codifica manuale, e di acquisire una maggiore precisione riducendo le occorrenze di codifiche errate, se paragonata alla codifica automatizzata non supervisionata.*

*Un confronto tra la correzione automatizzata supervisionata e quella non supervisionata è stato, infine, condotto per valutare quanto fossero distanti i risultati ottenuti dai due approcci metodologici.*

## 1. Introduction

The assessment is a fundamental phase of educational process (Berry, 2003; Cucchiarelli *et al.*, 2000), as it allows to verify and evaluate the knowledge acquired by the student.

In a typical examination setting, this assessment implies an instructor or a grader who provides students with feedback on their answers to questions that are related to the subject matter. There are, however, certain situation in which an instructor is not available and yet students need an assessment of their knowledge of the subject (Mohler and Mihalcea, 2009). In this situation computer based test are frequently used.

In Italy, the d.lgs. 62/2017 has introduced new norms in matter of evaluation and certification of the competences in the Italian school system. One of the most important innovations concerns the introduction of standardized Computer Based Tests (CBT) in national learning surveys for students in grades 8 (third upper secondary level), 10 (second secondary level) and 13 (fifth secondary level).



ISBN 9788835131557

CBT consists of an evaluation test in computerized form, no longer carried out with the use of paper and pen but with monitor and keyboard. In fact, nowadays the great technological innovations have allowed an ever greater implementation of this modality also in the educational field: at international level OECD has officially introduced the CBT mode for learning tests since the PISA 2015 survey, IEA with the 2016 ePIRLS surveys and TIMMS 2019.

Starting from s.y. 2017/2018, INVALSI has made, for the classes mentioned by d.lgs. 62/2017, the transition from a paper based test to a CBT test with the help of the computer. This transition has made it possible to use an increasingly wide range of types of questions to investigate students' competences more thoroughly (Scheuermann and Björnsson, 2009; Valenti *et al.*, 2000; Parshall *et al.*, 2000). However, the introduction of the CBT and the implementation of increasingly sophisticated and complex items inevitably puts the focus on automatic evaluation of tests (Automatic Assessment), and consequently on automated correction, or coding, of students responses.

Understandably, if multiple choice questions are easier to assess with an automated procedure and computational methods, open-ended questions (open-ended short answers) require a technology capable of evaluating natural language or structured text of mathematical notation (Burrows *et al.*, 2014). For this type of item, in fact, the number or type of characters – or words – that can be typed by the student (for example numbers and/or special characters) are not always constrained a priori in order to let the student free to express himself by simulating as much as possible the approach of the paper test.

Therefore, it is clear that in the educational context a correct assessment of the open-ended questions represents not only an essential objective but also, and above all, a great challenge. Trying to combine reliability, accuracy, impartiality and homogeneity in the correction of tests, the INVALSI team responsible for the correction of open-ended questions, adopted a hybrid approach: an automated supervised correction.

During the administration of the tests each student uses a computer connected to a main server for the collection of all the individual students' test, for all the subjects, in a timeframe defined by INVALSI: the set of Italian Language, Mathematics and English Language assessment constitute the database object of correction and coding by the INVALSI team. This research describes the new automated correction procedure introduced from 2018-2019 s.y. for open-ended questions in the INVALSI tests administered in CBT mode.

A correction approach fully automated oriented to "machine learning" is also presented to evaluate potentiality of coding of the same database.



ISBN 9788835131557

Moreover, will be illustrated in detail the phases of both procedures and will be explained the main results of the comparison between the codifications obtained from both procedures in order to evaluate the advantages and criticalities of the two different methodological approaches.

## 2. Data and methods

In Italy the d.lgs. 62/2017[1] has introduced important modifications for the INVALSI tests, starting from school year 2017/2018: tests are in fact administered to all students (census tests) of third upper secondary level, second and fifth secondary level schools, in CBT (Computer Based Test) mode, for Italian Language, Mathematics and English Language (Listening and Reading).

The new way of administering the tests prompt a new approach to coding the responses given by students to open-ended questions. The centralization of coding and the development of an ad hoc correction procedure with a certain degree of automation was necessary: maintaining an acceptable degree of accuracy and precision in coding, this procedure allowed the hours-limits to be overcome.

The degree of automation implemented in a correction procedure can intuitively be considered inversely related to the time-work requirements and the precision expected for the procedure itself. Let's consider on one side the classic correction or "manual", which maximizes the expected accuracy but also the need for hours-work, and on the other side the totally automatic correction carried out by a specific algorithm, which allows to minimize hours-work but predictably results in a greater frequency of correction errors, the procedure implemented in 2018 by the INVALSI team, that we could define as a "supervised automatic", achieves a good compromise in the trade-off between precision and work-time requirements.

Below we shown the description of the INVALSI's team procedure applied on Italian Language and Mathematics tests administrated in CBT mode in third upper secondary level school (below grade 8$^{th}$) of school year 2018-19. Then a second procedure is described following the "machine learning" approach (ML): a totally automated procedure able to minimize the need for hours-work. Considering the "automatic supervised" procedure the gold standard to refer to, an analysis of the coding performance of two different ML algorithms was carried out in order to frame and compare the trade-off

[1] *Gazzetta Ufficiale della Repubblica Italiana*, Decreto legislativo 13 aprile 2017, n. 62, retrieved on April 6, 2021, from: https://www.gazzettaufficiale.it/eli/id/2017/05/16/17G00070/sg.



ISBN 9788835131557

between precision and hours-work requirements for the two procedures. For the comparison between the two algorithms ML has been selected from the database a small set of 10 items of Italian Language test.

### *2.1. The new INVALSI automatic supervised procedure*

The INVALSI team, consisting of statisticians and computer scientists, has implemented a complex procedure of coding of the answers provided by the students, obtaining an algorithm for the treatment of more or less articulated text strings. This procedure involves carrying out some important preparatory operations which we could define as the "testing phase":

– for each subject (Italian Language, Mathematics, English Language) groups of authors define with the correction team the rules or the correction criteria in a clear and unambiguous way, according which it is possible to discriminate the correct answers from the wrong ones in reference to each item;
– the correction criteria were translated in to computer logic patterns (more in detail called *regular expressions* or "RegEx") which can be interpreted by the computer and are useful to classify the processed answers in a completely automatic way;
– the correction team performs the actual testing of the algorithm by verifying, for the students answers during the pre-test phase, the degree of concordance between the coding produced by manual correction and the other processed automatically by the algorithm. The algorithm and the logic-computer patterns that translate the correction conditions are considered sufficiently accurate and aligned to the authors' coding indications when there is perfect convergence between the two coding;
– the correction team plans together with the authors a series of control activities and the production of reports to be carried out during the period of administration of the tests. The purpose of these activities is to verify the accuracy of the coding process and, if necessary in authors' opinion, to modify the conditions.

After the test phase, the set of operations which encode the open-ended question consists in four distinct steps:

1) *acquisition of database* containing the students answers to open-ended questions and referred to a defined time frame of administration;
2) *data cleaning operations*: the answers are subjected to automatic correction of typing and/or spelling errors and are deprived of all the elements considered not useful for correction (punctuation marks, parentheses, etc.);



ISBN 9788835131557

3) *encoding operations*: the algorithm compares each response with the logical version of the correction conditions (the regular expressions or RegEx) classifying it as correct or incorrect;
4) *production of report* to provide a clear summary of the algorithm's responses classification.

In the first phase of the procedure, the database is acquired and targeted checks are carried out to highlight any macro anomalies in the data and in their structure.

In the data cleaning phase starts the real manipulation of the responses. This phase consists of a series of operations on the text strings in order to eliminate the invariant elements for the classification, reducing the variability of the possible modalities of response and, therefore, the complexity of the set to be classified. More in detail, the operations generally applied to the text strings include:

– identification and removal of punctuation, special characters, proposition and conjunctions;
– lemmatization of words, namely the reduction of an inflected form of a word to its canonical form;
– automatic correction of typing and spelling errors or detection of words "out of vocabulary" and replacement of words with the corresponding corrected version.

After data cleaning process, the "normalized" dataset is processed by an algorithm that performs the classification of each string in correct or incorrect using a "RegEx" engine. This software is capable of comparing text strings with the logical version of correction criteria called "regular expression" or "RegEx", refined in pre-test phase. A regular expression is a sequence of characters that uniquely identifies even an unfinished set of strings. Translate a correction condition into regular expression means defining the possible set of correct answers to a specific item, therefore the regex engine processes the strings of answer and classifies as correct only those that belong to the set previously defined from the RegEx.

The final phase involves the production of correction reports at intervals agreed with the authors, throughout the administrating window. The purpose of the reports is to provide a tool to assess the coding quality of the algorithm and to highlight any classification errors. Each report provides the frequency distribution of the students' responses and their assigned classification, for each time interval considered. The main addressees of the reports are the authors which are allowed to verify if it is necessary to introduce changes to correction conditions to improve the accuracy and efficacy of the classification.



ISBN 9788835131557

All the operations of the coding procedure described above have been implemented through the open source KNIME Analytics Platform (Berthold *et al.*, 2008).

## *2.2. Machine learning algorithms and fully automated correction procedure*

A typical automatic classification system for short answers (ASAG) focuses on assessing short natural language responses to questions in an automatic way, and automatically classify student answer into, correct or incorrect, based on the resemblance to referred correct one(s). In the ASAG field, in all disciplines but more often for the correction of short answer like understanding a text (Vijaymeena *et al.*, 2016; Meurers *et al.*, 2011), the methods of Machine Learning (ML) represent one of the approaches more used in literature (Burrows *et al.*, 2015; Galhardi and Brancher, 2018). In a machine learning system, classification is a supervised learning category.

The algorithm is trained by a supervisor to recognize categories (e.g. correct/incorrect) through a series of practical examples (dataset training); in each of the examples, the machine is provided with the descriptive variables of the working environment and a label to indicate the desired result. The system elaborates the examples searching a general rule of classification, defined model. When the model is obtained, the machine uses it to classify the new instances, making use of the observations on the training set.

Choosing the best machine learning algorithm for classifying student responses mainly takes into account the nature of the dataset, namely the number of records, the subject of the questions given, the average length of the answers, and finally the rating scale, which can be represented by a simple classification or a numerical score (Galhardi and Brancher, 2018).

Regarding the evaluation of the average length of student responses, a short string of text is often defined in literature between 5 and 20 words (Kitchenham, 2004).

In a preliminary phase of the algorithm, called pre-processing, the words contained in the students' answers are filtered and manipulated so that only the relevant words in the text can be considered during the correction process. The main techniques used in the pre-processing phase are:

– the removal of punctuation and special characters;
– the removal of *stop-words*, words that due to their high frequency in a language, are usually considered insignificant. Examples of stop word can be articles, propositions or conjunctions;



ISBN 9788835131557

– correction of spelling;
– the *emblem*, that is the process of reducing words to their root;
– lemmatization, that is the process of reduction in the inflected form of a word to its canonical form, called the lemma.

After the pre-processing phase, we choose the best classification algorithm, using, for the specific case, the family of supervised automated classification algorithms, as every answer in the dataset has already been classified as correct or wrong. From the answers already classified, the algorithm can learn the basic rules of correction and in this way it is able to classify a new response as correct or wrong.

In this work we evaluated the performance of two of the most used machine learning algorithms applied to the automated classification of short answer: the *Support Vector Machine* and the *Decision Tree.*

The Support Vector Machine (SVM) (Keerthi *et al.*, 2001) is a supervised machine learning algorithm generally used in binary classification problems. The idea behind the algorithm is to find a hyperplane that best divides a data set into two classes. Support vectors (Support Vector) are the data points closest to the hyperplane. Among all possible hyperplanes, the algorithm determines the one able to separate classes with as much margin as possible. In this context, the margin can be defined as the distance between the support vectors of two different classes closer to the hyperplane. Therefore, a good separation is obtained from the hyperplane which has the greatest distance from the supporting vectors, that is from the nearest points of each of the two classes. In general, the greater the margin, the less generalization error is made by the algorithm.

In the artificial intelligence field, a *Decision Tree* (Quinlan, 1993; Morent *et al.*, 2011), is a model that uses a tree-shaped data structure to contain information and make predictions. Their nature is such that they turn out to be models more interpretable than others, in fact they express in the arcs and in the nodes the conditions that generate the prediction. The training of the model will consist in reaching the leaves of the tree using as few conditions as possible. The set of potential binary cuts that divide the units contained in a parent node into two sets that form child nodes is called split set. The most commonly used parameter for split conditions is the *Gini Index*, which reaches its minimum (zero) when the node belongs to a single category.

The first step for the performances comparison of the two algorithms consists in the creation of a *bag of word*, that is a list of words contained in each answer. Starting from this list, is created a vector whose elements are numbers associated to each of the words in the dataset, after the pre-processing phase.



ISBN 9788835131557

These values are obtained using the *Inverse Document Frequency* (IDF) function, which allocate high value to specific words and low value to generic words. This function measures the importance of a specific word or term in the text or in an entire document. Some words, indeed, that are frequently in a text could be not relevant. The importance of the term in the text can be calculated with the formula:

$$IDF\,(i) = log\ \frac{N}{N_i}$$

where N is the total of texts or documents analyzed,  is the number of documents that contain the *i* term. Generally, this indicator is multiplied by the *Term Frequency* (TF) which measures the number of times a word appears in a document. Clearly, TF is more important in longer texts, such as essays, but it is not a relevant indicator in the case of short answers.

After obtaining the vectors, the dataset is divided into two categories:

– the *training set*, that is the dataset used to train the algorithm;
– the *test set*, that is the dataset used to evaluate the performance of the algorithm.

Finally, the degree of accuracy and reliability of the classification, obtained with the SVM algorithms and *Decision Tree*, was measured using the *Cohen's Kappa coefficient*.

*Cohen's coefficient K* (Cohen, 1968) measures the degree of accuracy of a statistical classification, obtained by comparing the observed agreement with the random agreement. This is an index of concordance calculated on the basis of the relationship between the agreement in excess of the probability of random concordance and the maximum obtainable excess.

Starting from the confusion matrix, the index can be calculated with the formula:

$$k = \frac{p_0 - p_e}{1 - p_e}$$

where $p_0$ represents the percentage of agreed valuations and it is equal to the sum of the first diagonal of the matrix divided by the total of the valuations; $p_e$ represents the probability of random agreement and consists of the product of the positive totals plus the negative totals, all divided by the square of total valuations.



ISBN 9788835131557

## 3. Results

The core of the coding procedure implemented by INVALSI and described above is clearly the set of regular expressions associated with the group of items constituting a test. A regular expression must be able to intercept, for each item, both the correct answer and also all the more probable variants of it considered correct by the authors.

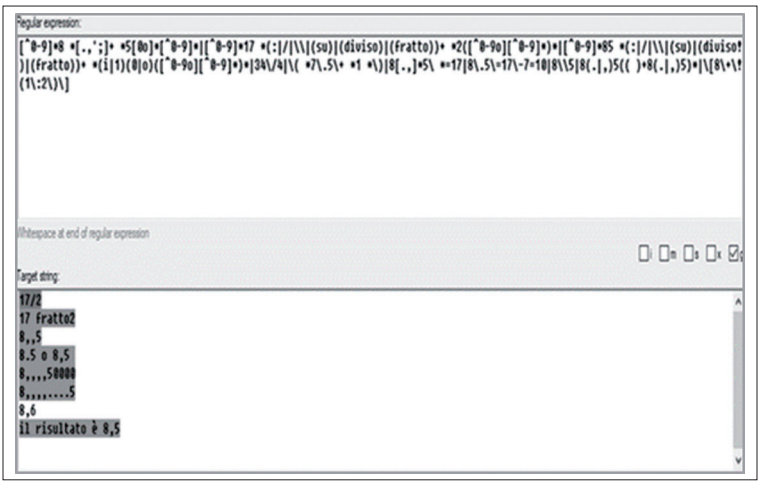


*Fig. 1 – An example of translation of a correction condition into regular expression*

For example, in Fig. 1 is presented an example of RegEx (in the upper box) which, by construction, can identify (in the lower box) not only the correct "official" answer (in this case 17/2) but also a whole series of probably correct variants of it (8.5; 8.5; 17 divided by 2). The most complex aspect in the construction of regular expressions is precisely to endow them with an acceptable discriminatory sensitivity.

Quality and flexibility of each regular expression determine the overall quality of the encoding produced: a set of inflexible regular expressions would significantly increase the number of false negatives, that is all variants to the correct response key not recognized by the RegEx. Considering the high number of codified tests during a national survey, the possibility that a large number of variants of the correction keys will occur in the student's responses is a rather likely scenario; to confirm this thesis, are indicative the tables below (Tab. 1 and 2) from which it can be deduced the high number of variants of the correction key that the implemented RegEx have managed to codify correctly.



ISBN 9788835131557

*Tab. 1 – Italian Language test, Main Study 2019, grade 8th*

| *Item* | *Interaction type* | *No. variants of correction key* |
|---|---|---|
| Item 1 | Extended text | 3,942 |
| Item 2 | Extended text | 3,289 |
| Item 3 | Extended text | 1,810 |
| Item 4 | Text entry | 1,619 |
| Item 5 | Block + Text entry | 1,298 |
| Item 6 | Extended text | 1,175 |
| Item 7 | Block + Text entry | 927 |
| Item 8 | Text entry | 714 |
| Item 9 | Extended text | 673 |
| Item 10 | Extended text | 541 |
| Item 11 | Block + Text entry | 492 |
| Item 12 | Block + Text entry | 342 |
| Item 13 | Text entry | 291 |
| Item 14 | Extended text | 286 |
| Item 15 | Extended text | 256 |
| Item 16 | Block + Text entry | 232 |

*Tab. 2 – Mathematics test, Main Study 2019, grade 8th*

| *Item* | *Interaction type* | *No. variants of correction key* |
|---|---|---|
| Item 1 | Text entry | 650 |
| Item 2 | Text entry | 381 |
| Item 3 | Block + Text entry | 375 |
| Item 4 | Text entry | 350 |
| Item 5 | Block + Text entry | 322 |
| Item 6 | Text entry | 317 |
| Item 7 | Text entry | 306 |
| Item 8 | Block + Text entry | 184 |
| Item 9 | Block + Text entry | 155 |
| Item 10 | Text entry | 119 |
| Item 11 | Block + Text entry | 112 |
| Item 12 | Text entry | 96 |
| Item 13 | Text entry | 94 |
| Item 14 | Block + Text entry | 93 |
| Item 15 | Text entry | 87 |
| Item 16 | Text entry | 76 |
| Item 17 | Block + Text entry | 75 |
| Item 18 | Block + Text entry | 73 |



ISBN 9788835131557

For a comparison between the performance of the two ML algorithms and, subsequently, for evaluation of the degree of agreement between the ML coding and the gold standard of the "assisted" INVALSI procedure, 10 items of national survey in Italian Language (grade 8th) were selected, whose answers count an average number of words between 5 and 7. For each item the algorithm has acquired information from about 100 thousand answers already classified in correct/incorrect, which constitute the dataset of analysis. The dataset was divided so that 70% of it was used as a training set and 30% as a test set.

The values of the *Kappa coefficient* obtained are quite similar between the two algorithms (Tab. 3), although the Support Vector Machine seems to provide better results in cases of more complex classification. Many of the errors traced are attributable to cases of misspelling that the automatic corrector was unable to identify and correct. For example, some cases have been identified in which two or more words have been written without separator or with typing or spelling errors.

*Tab. 3 – Performance comparison: SVM and Decision Tree algorithm. Italian Language tests, grade 8th*

| | *SVM* | | *Decision Tree* | |
|---|---|---|---|---|
| | *K-score* | *False negative + False positive* | *K-score* | *False negative + False positive* |
| Item 1 | 0.98104 | 1,182 | 0.98042 | 1,176 |
| Item 2 | 0.99293 | 283 | 0.99549 | 203 |
| Item 3 | 0.99452 | 42 | 0.99564 | 37 |
| Item 4 | 0.99623 | 71 | 0.99566 | 68 |
| Item 5 | 0.99148 | 247 | 0.99623 | 126 |
| Item 6 | 0.99613 | 97 | 0.99648 | 90 |
| Item 7 | 0.99484 | 255 | 0.99674 | 172 |
| Item 8 | 0.99802 | 76 | 0.99792 | 79 |
| Item 9 | 0.99881 | 12 | 0.99891 | 11 |
| Item 10 | 0.99843 | 16 | 0.99902 | 10 |

Anyway, being able to count, for each item, on about 100 thousand already classified answers, the values of K statistic are always higher than 0.98, for both algorithms, therefore very close to the maximum 1. However, even if with very high values of Kappa, in some cases, the algorithm makes hundreds of classification errors.

Taking in high consideration the fact that, in the case of correction of the open-ended questions from the INVALSI tests, errors in the classification by the algorithm would determinate the attribution to the student of a smaller



ISBN 9788835131557

number of correct answers than the actual one, therefore to an underestimation of its level in the competences and consequently the delivery of an erroneous certification.

## 4. Conclusion

The introduction in the school system of standardized Computer-based tests for the survey of students' competences has made indispensable the adaptation of an automated approach in the correction procedure. The automated supervised correction method adopted by INVALSI for the open-ended questions required, in the different stages of implementation, a close comparison between correction team, consisting of statisticians and computer scientists, and the groups of authors of each subject (Italian Language, Mathematics, English Language). This long and complex work, however, has allowed to obtain an algorithm that guarantees, with an accuracy of almost 100%, a correct classification of student responses. The ultimate goal of automatic coding, in fact, is to issue schools and students with a correct certification of competences, free from classification errors.

It is clear, however, that comparing to a fully automated correction procedure, such as that used with machine learning algorithms, the automated supervised procedure requires more hours of work and involves more resources.

Anyway the adoption of good practice, the improvement of pre-processing procedures and the creation of a database containing a collection of thousands of types of student responses, could lead to a significant reduction in the amount of working hours.

The method adopted by the INVALSI correction team can therefore represent a good compromise between manual coding and the adoption of machine learning algorithms to achieve performance that maximize coding accuracy.

On the other hand, it would be advisable pursue in the future a reflection between correction teams and expert groups of different disciplines to evaluate the adoption of fully automated models which, if well designed and prepared, may provide good guarantees for timely, effective and accurate performance.



ISBN 9788835131557

## References


Berry R. (2003), "Alternative assessment and assessment for learning", in *Proceedings of the 29th IAEA Conference, "Societies Goals and Assessment", USA*.

Berthold M.R., Cebron N., Dill F., Gabriel T.R., Kotter T. (2008), "*KNIME: The Konstanz Information Miner*", in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (*eds.*), *"Data Analysis, Machine Learning and Applications. Studies in Classification", Data Analysis, and Knowledge Organization*, Springer, Berlin, Heidelberg.

Burrows S., Gurevych I., Stein B. (2014), "The eras and trends of automatic short answer grading", *Int. J. Artif. Intell. Educ.*, 25, pp. 60-117.

Cohen J. (1968), "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit", *Psychological Bulletin*, 70, 4, pp. 213-220.

Cucchiarelli A., Panti M., Valenti S. (2000), "Web-based assessment of Student Learning", in A.K. Aaggarwal (*ed.*), *Web-Based Learning and Teaching Technologies: Opportunities and Challenges*, Idea Group Publishing, retrieved on July 8, 2021, from: http://www.jite.informingscience.org/documents/Vol1/v1n3p157-175.pdf, pp. 175-197.

De Mauro A. (2019), *Big data analytics. Analizzare e interpretare dati con il machine learning*, Apogeo, Milano.

Galhardi L.B., Brancher J.D. (2018), "Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review", in G. Simari, E. Fermé, F. Gutiérrez Segura, J. Rodríguez Melquiades (*eds.*), *Advances in Artificial Intelligence. IBERAMIA 2018. Lecture Notes in Computer Science*, Springer, Cham.

INVALSI (2019), *Le prove Computer Based (CBT). Terzo anno scuola secondaria di I grado (grado 8) a.s. 2019-2020. Organizzazione delle prove CBT*, retrieved on March 8, 2021, from: https://INVALSI-areaprove.cineca.it/docs/2020/2019-2020_Organizzazione%20delle%20prove%20CBT_Grado_08.pdf.

Keerthi S.S., Shevade S.K., Bhattacharyya C., Murthy K.R.K. (2001), "Improvements to Platt's SMO algorithm for SVM classifier design", *Neural Computation*, 13, pp. 637-649.

Magliano J.P., Graesser A.C. (2012), "Computer-based assessment of student-constructed responses", *Behavior Research Methods*, 44, 3, pp. 608-621.

Meurers D., Ziai R., Ott N., Kopp J. (2011), "Evaluating answers to reading comprehension questions in context: results for German and the role of information structure", in *Proceedings of the TextInfer 2011, Workshop on Textual Entailment*, pp. 1-9.

Minini A., *Gli alberi di decisione*, retrieved on March 8, 2021, from: http://www.andreaminini.com/ai/machine-learning/alberi-di-decisione.

Molher M., Mihalcea R. (2009), "Text-to-Text Semantic Similarity for Automatic Short Answer Grading", *Conference: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, March 30-April 3, 2009, Athens, Greece*.




ISBN 9788835131557

Morent D., Stathatos K., Lin W.C., Berthold M.R. (2011), "Comprehensive PMML preprocessing in KNIME", in *Proceedings of the 2011 workshop on predictive markup language modelling*, ACM, San Diego (CA).

Parshall C.G., Davey T., Pashley P.J. (2000), "Innovative Item Types for Computerized Testing", in W.J. van der Linden, G.A. Glas (*eds.*), *Computerized Adaptive Testing: Theory and Practice*, Springer, Dordrecht (NL).

Quinlan J.R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Los Altos (CA).

Ray S. (2017), "Understanding Support Vector Machine(SVM) algorithm from examples (along with code)", *Analytics Vidhya*, retrieved on March 8, 2021, from: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, September 13.

Scheuermann F., Björnsson J. (2009), *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, Office for Official Publications of the European Communitites, Luxembourg.

Vijaymeena M., Kavitha K. (2016), "A survey on similarity measures in text mining", *Mach. Learn. Appl.: Int. J.*, 3, 2, pp. 19-28.



ISBN 9788835131557

ISBN 9788835131557

# *The authors*

**Francesco Annunziata**, graduate in Sociology and Policy-Making from a Territorial Perspective at the University of Salerno, works in the International Surveys department at INVALSI monitoring test administrations and verifying the consistency of data for OECD and IEA surveys.

**Cecilia Bagnarol**, graduated in Statistics, Economics and Business at the Alma Mater Studiorum of Bologna. She currently works at Statistical Service of INVALSI where she performs support activities for statistical analysis on large data bases of national surveys on learning.

**Giorgio Bolondi** is a Mathematician, Phd in Algebraic Geometry, is interested in how mathematical knowledge goes from generation to generation and from person to person. He teaches at the Free University of Bozen/Bolzano; his current research activity is focused on the assessment of learning and the professional development of Mathematics teachers.

**Emiliano Campodifiori**, graduated in Statistics and Economics at the University of Rome "La Sapienza". Currently he works in the Statistical Service of INVALSI, he performs statistical analysis of the National Assessment data.

**Elisa Caponera** is a researcher at INVALSI. She was italian National Research Coordinator (NRC) for TIMSS 2011 and ICILS 2018 projects. Her current themes of research are parent involvement at school, gender difference in Mathematics achievement, school effectiveness and equity of school system.



ISBN 9788835131557

**Simone Del Sarto**, post-doc researcher at the Department of Statistics, Computer science, Applications "Giuseppe Parenti" of the University of Florence. He earned the PhD in Statistics in 2015 at the University of Perugia. His main interests deal with latent variable models (specifically, Item Response Theory, latent class and latent Markov models) and composite indicators construction in the field of education and corruption measurement.

**Silvia Donno**, graduated in Demographic Sciences for Social and Health Policies at the University of Rome "La Sapienza". Currently she works in Statistical Service of INVALSI, she carries out activities to support the elaboration and statistical analysis of the data of the national surveys on learning.

**Federica Ferretti**, PhD in Mathematics, researcher in Mathematics Education at Department of Mathematics and Computer Science, University of Ferrara. Her research concerns the Didactic Contract at all school levels, formative assessment in Mathematics and the formative use of standardized assessment. For years she has been involved in Mathematics teacher's professional development.

**Michela Gnaldi**, associate Professor at the Department of Political Science, University of Perugia. She has been scientific responsible of two research agreements with INVALSI. She is co-author of many scientific papers, including "Students' Complex Problem Solving Profiles" (Psychometrika, 2020) and the book "Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata" (Chapman and Hall, 2016).

**Maria Magdalena Isac** is a researcher at KU Leuven, Belgium and INVALSI, Italy. Her research is focused in the area of comparative evaluation of educational systems, with emphasis on understanding how different educational approaches contribute to young people's citizenship learning and on the use of large-scale assessment data.

**Michele Marsili**, graduated in Statistics at Sapienza University of Rome. He worked in Business Intelligence consulting, providing software development solutions for analysis and support for company's decision making in insurance and pharmaceutical industries. Since January 2018 he has been working in the Statistical Service of INVALSI.

**Laura Palmerio**, senior researcher at INVALSI, is head of the International Surveys department. Italian Responsible for OECD and IEA projects.



ISBN 9788835131557

She is a member of INVALSI Scientific Council and of the TIMSS Questionnaire Item Review Committee. Main research interests: equity in education, relations between literacy in reading and in Mathematics.

**George Santi**, PhD in Mathematics, is a researcher at the University of Bolzano. His research focuses on the networking of semiotic perspectives in Mathematics Education related to several research issues concerning the teaching and learning of Mathematics.

**Zuzana Toth** is research fellow at INVALSI. Her research focuses on the development of linguistic competence and language awareness in L1, L2 and L3(s).



ISBN 9788835131557

# Vi aspettiamo su:

**www.francoangeli.it**

**per scaricare (gratuitamente) i cataloghi delle nostre pubblicazioni**

DIVISI PER ARGOMENTI E CENTINAIA DI VOCI: PER FACILITARE
LE VOSTRE RICERCHE.


**Management, finanza, marketing, operations, HR**
**Psicologia e psicoterapia: teorie e tecniche**
**Didattica, scienze della formazione**
**Economia, economia aziendale**
**Sociologia**
**Antropologia**
**Comunicazione e media**
**Medicina, sanità**



**Architettura, design, territorio**
**Informatica, ingegneria**
**Scienze**
**Filosofia, letteratura, linguistica, storia**
**Politica, diritto**
**Psicologia, benessere, autoaiuto**
**Efficacia personale**
**Politiche e servizi sociali**

FrancoAngeli
La passione per le conoscenze
ISBN 9788835131557



**Comunicaci il tuo giudizio su:**
www.francoangeli.it/latuaopinione.asp

VUOI RICEVERE GLI AGGIORNAMENTI
SULLE NOSTRE NOVITÀ
NELLE AREE CHE TI INTERESSANO?


ISCRIVITI ALLE NOSTRE NEWSLETTER SEGUICI SU:

FrancoAngeli
La passione per le conoscenze
ISBN 9788835131557

The school system has always aimed to achieve quality teaching, which is able, on the one hand, to give adequate responses to the expectations of all the stakeholders and, on the other, to introduce tools, actions, and checks through which the training offer can be constantly improved. This process is undoubtedly linked to scientific research. Researchers and Academics start from the data available to them or collect new ones, to discover and/or interpret facts and to find answers and new cues of reflection. A favorable environment for this work was the Seminar "INVALSI data: a research and educational teaching tool", in its fourth edition in November 2019. The volume consists of six chapters, which are arise within the aforementioned Seminar context and, while dealing with heterogeneous topics, offer important examples of research both on teaching and on the methodologies applied to it. As a Statistical Service, which for years has taken care of the collection and dissemination of data, we hope that in this, as in the other volumes of the series, the reader will find confirmation of the importance that data play, both in scientific research and in practice in classroom.

**Patrizia Falzetti** is Head of the INVALSI Statistical Service, which manages the acquisition, analysis and return of data concerning national and international surveys on learning to individual schools, stakeholders and the scientific community.

FrancoAngeli
La passione per le conoscenze