Studies in Educational Evaluation 58 (2018) 37–50

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

# A quantitative methodology for analyzing the impact of the formulation of a mathematical item on students learning assessment

Giorgio Bolondi[a], Laura Branchetti[b,*], Chiara Giberti[c]

[a] Department of Education, Free University of Bolzano, Viale Ratisbona, 16, 39042, Bressanone-Brixen (BZ), Italy
[b] Department of Mathematics, Physics and Computer science, University of Parma, Parco Area delle scienze, 7/a, 43124, Parma (PR), Italy
[c] I.C. "Fabriani", Viale Marconi, 6, 41057, Spilamberto, Modena (MO), Italy

ARTICLE INFO

*Keywords:*
Student evaluation
Evaluation methods
Word problems text formulation
Item response theory

ABSTRACT

In this paper, we present a methodological approach to the investigation of the impacts of text formulation on students' answers in mathematical problem solving-based assessment. After a review of the related literature in Mathematics education and a review of the methodologies used until now to investigate this research issue, we describe in depth our quantitative approach, providing motivations and examples of its statistical relevance and its potentiality in making interesting phenomena emerge, to be interpreted with further qualitative methods. We observed statistically significant evidences of different impacts of the variations on different categories of students (males/females; students with high and low performances in the whole test). The methodology and our preliminary results can inform researchers in mathematics education, teachers and experts in the agencies that are responsible for large-scale students learning assessment in several contexts (national and international).

## 1. Introduction

When facing a mathematical task, students are influenced by the formulation of the task itself. In particular, this may influence in a significant way their performance when dealing with an assessment task, for instance in the case of a word problem. This is a classical topic in educational research; for instance, a recent literature review for the case of arithmetical word problems is Daroczy, Wolska, Meurers, and Nuerk (2015).

A better understanding of the relationships between formulation of a problem/task, reading and problem-solving strategies and students' performances may have three kinds of impact:

- a *theoretical* one, in the direction of problematizing the relation between students' knowledge and the assessment based on students' answers to written tests as "final products": it can contribute to better define the summative aspects of assessment;
- a *practical* one: it may help task-assessment designers (teachers, large-scale assessment authors, researchers….) both in well defining the *question intent* and in monitoring different levels of difficulty;
- a *didactical* one: it may help in interpreting students' behaviors when answering to an assessment question, hence, it can also give a contribution to formative assessment.

In this paper, we propose our methodological contribution to this general problem by designing and validating a quantitative methodology for measuring the impact of a variation in the formulation of an item on students' performances. In particular we present: our background; the steps and the kind of data necessary to carry out a research based on this methodology; a validation plan of the methodology based on the confirmation in two cases of results that we consider "solid findings" in Mathematics education (Education Committee of the EMS, 2011) concerning the impact of formulation in mathematical problem-solving; two examples of quantitative results that put new light on those findings and may encourage researchers in Mathematics education to carry out further qualitative researches on new categories of phenomena that have not been investigated yet since the methodologies used to address the research questions did not take care of such aspects.

The structure of the paper is as follows.

In Section 2, we outline the background of the problem. We present in §2.1 different approaches to the categorization of the variables in the formulation of a mathematical task and some research on the impact of different formulations on the performance of students. This review will allow us in §4 to frame the cases that we will use for the validation of our tool. In §2.2 we review the different methodologies used in the research on the impact of variation of formulation of a task, presenting the main methodological difficulty and showing the lack of a quantitative method for measuring this impact, hence the rationale for this

* Corresponding author.
*E-mail addresses:* giorgio.bolondi@unibz.it (G. Bolondi), laura.branchetti@unipr.it (L. Branchetti), chiara.giberti@unitn.it (C. Giberti).

https://doi.org/10.1016/j.stueduc.2018.05.002
Received 9 January 2018; Received in revised form 22 April 2018; Accepted 11 May 2018
Available online 26 May 2018
0191-491X/ © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

paper. In §2.3 we present the statistical background of our methodology, which is indeed the set of techniques largely used in large scale assessments. §2.4 contains our research questions.

Section 3 is devoted to the description of our method. §3.1 contains the design of the tool, §3.2 the outputs, and §3.3 the coherence and compatibility conditions that must be satisfied in order to consider acceptable the data obtained.

Section 4 contains the validation plan and its results, which is based on a starting test (coming from a large-scale assessment) for which solid data are already available, and variations of formulation for which extensive didactic researches have been already performed. We stress since now the fact that our purpose is not to interpret data in order to provide new results at this stage. Our validation strategy relies on showing how our tool provides data both confirming previous results (obtained with different methodologies), and pointing out new phenomena. In §4.1 we describe our starting test, the variations, and the population. In §4.2 we verify that our test satisfies the coherence and compatibility conditions 3.3, and that our experiment provides general data coherent with the solid data of the original large-scale assessment. In §4.3 we present the output data for two cases and we discuss them under the light of existing didactic research, showing what our method may provide for a quantitative framing of the phenomena. In particular, we show how it highlight new phenomena.

Section 5 contains our conclusions, with our remarks about the future issues and the limit of our approach.

## 2. Background of the research and statement of the problem

### 2.1. Variables in the formulation of a problem

We present here the context where our analysis takes place. This review of previous researches will furnish the didactic variables for our study.

During the last decades, many authors studied and classified possible formulations of the test of a mathematical task. Others, inquired about the impact of differences in the test impact on students' behavior. We report in this section some relevant results that we used as solid findings to check with our methodology and, in particular, we describe two results that we analyzed in depths, as we report in the Section concerning the data analysis.

As pointed out by Bagni and D'Amore (2005), the crucial point concerning variations in the formulation is not the fact that a formulation is necessarily better or worse than another one, but the fact that changing the formulation actually changes the problem.

The factors influencing students' approach to the answering of a written test may be a lot and it is complex to list them being exhaustive. However, some attempts have been done to list categories of such factors in the field of mathematics education and we started from them to have a picture of what could be interesting to investigate while facing the problem of measuring the impact of variations in the formulation on students' performances. Analyzing the factors affecting the problem-solving activities, Nesher (1982), while categorizing the variations, listed three components that may vary in a *word problem*: logical (operations, lack or abundance of data, …), syntactic (position of the question in the text, number of words, …) and semantic (contextual relations, implicit suggestions, …).

Considering the more general problem (not necessarily bounded to a mathematical word problem or to arithmetical contents) of the comprehension of a text and the information retrieval, Duval (1991) studied what he called "variables rédactionnelles" (French original name), stating that they influence the student´s cognitive and operative processes. As D'Amore (2000) highlighted, these modifications in the text, even small, may cause changes in the students' approaches to problem solving. Laborde redefined them in 1995 in order to include also other non-verbal variations, such as introducing form of representations (Laborde, 1995). She listed factors concerning editing, punctuation, syntactical complexity, word density, order of the information, explicit declaration of intermediate objects needed for the solution. However, how individuals come up with mathematical solution strategies can also be influenced by numerical factors like number magnitude (Thevenot & Oakhill, 2005). This result is confirmed and analyzed in depths by De Corte, Verschaffel, and Van Coillie (1988), with a focus on the number type (integer, decimal bigger than 1, decimal smaller than 1) in arithmetical word problems concerning multiplication, stressing the difference in students' answers when the number type change in the multiplier, while they stressed that there were no significant changes when the multiplicand changed. We use this as first solid finding to analyse in depths to validate our methodology since it is very detailed and strong from the methodological point of view, ans it has been also mentioned as a solid finding by Daroczy et al. (2015) when they proposed a review of the factors affecting the difficulty of word problems and described the "three components of WP difficulty: (i) the linguistic complexity of the problem text itself, (ii) the numerical complexity of the arithmetic problem, and (iii) the relation between the linguistic and the numerical complexity of a problem". Yet, Daroczy et al. (2015) stated that variations in problem solving strategies could depend on linguistic factors like wording, semantic categories and propositions. The influence of linguistic factors on Mathematics teaching and learning is a classical topic in Mathematics education – see for instance the review by Schleppegrell (2007), that has been investigated a lot in the case of problem solving. A detailed analysis of *word problems* that is relevant from this point of view has been carried out by Frank, Koppen, Noordman, Vonk, and Perfetti (2007). According to the authors (p. 2): "A broad model of text comprehension should not only simulate how information is extracted from the text itself, but also how this information is interpreted in light of the reader's knowledge." This distinction is related to the distinction among three levels of discourse representation: the first level is the *surface representation*, "consisting of the text's literal wording"; the second level "called the *textbase*, where the meaning of the text is represented as a network of concepts and propositions from the text […] connection relations between propositions in a coherent text base are typically expressed by connectives"; the third level of representation, named *situation model*, "textbase elements are combined with elements from the reader's general knowledge".

Branchetti and Viale (2015) contributed to the general statement that linguistic variations can affect students' performances. In particular, they investigated the impact of variations in the syntactic structure of the sentences (variating thus the first and the second level) and highlighted effects of these variations also on students with good performances in mathematics. We referred to the general solid findings concerning the impact of variations of the linguistic factors reported by Daroczy et al. (2015), trying to investigate from a quantitative point of view the statement by Branchetti and Viale (2015) about the students with good performances, comparing two cases that our methodology showed to be very different from the point of view of the students answers distribution

### 2.2. Research methodologies

The methodologies used in order to investigate the impact of these variations are almost quantitative and often consist in the administration of different tests, containing two or more formulations of the "same" task. In some studies, the same question is revised and reformulated in many versions and all the different forms of the task are administered to the same group of students (e.g. Lepik, 1990; Cummins, Kintsch, Reusser, & Weimer, 1988; De Corte, Verschaffel, & De Win, 1985; Thevenot, Devidal, Barrouillet, & Fayol, 2007). In this case, the ability of the students responding to the different versions of each task is the same but the main problem of this approach consists in the unavoidable influence of the work performed by the student on the first task administered, on his resolution of the second one. In almost all of these researches, the way to partially overcome this obstacle consists in

changing the order of the versions proposed to the students or to allow time to pass between when the student faces the first version and when he faces the other version (e.g. Vicente, Orrantia, & Verschaffel, 2007). For example, in the research of De Corte et al. (1985) the authors prepared two tests: in test A the problems appear in a form similar to those of textbooks problems, and in test B the same problems were reformulated to be more clear to students. They administered both of the tests to a sample of 170 students in two sessions, one week apart, but half of the students faced first test A, and one-week later test B, whilst the other half faced the two test in the opposite order.

Other quantitative approaches to this issue try to overcome the obstacle explained before by using different populations of students, and this is the case for instance of Nesher research (1976): 4 different tests containing different versions of the same problems were administered to 800 students in total, but each student answered only to one version of the test.

In some cases, we find also qualitative researches having the goal of analyzing the impact of a variation, based on interviews of the students (where they were asked to compare the two versions of a task) or based on the analysis of protocols of the students (e.g. Spanos, Rhodes, & Dale, 1988).

In general, research methodologies can be framed either in quantitative methods or in qualitative methods. Nevertheless, recent educational research is moving more and more towards a mixed method approach (Johnson & Onwuegbuzie, 2004). In this direction we found the work of Abedi and Lord (2001) that combine two steps: the first one consists in interviews of students and the second one is quantitative and based on two versions of the same test administered to a sample of 1174 students. Moreover, this research is particularly interesting for us because the two tests where composed by 20 varied problems (which were presented in the original form in one of the two tests and in the varied form in the other) and 5 control items, unchanged in the two tests.

Branchetti and Viale (2015) presented another example of mixed method research on this issue. In particular, they proposed a methodology based on the IRT (*Item Response Theory*) and on the Rasch model (Rasch, 1960) in order to study the problem of the syntactic structure of the formulation of a task. They carried out a pilot study with about 200 students concerning linguistic changes in the text formulation. They changed the syntax of some problems of a large-scale standardized test and asked the students to answer the whole modified test (including the not modified questions). The authors decided to compare the expected students' answers to the items formulated in the original way (prediction based on a statistical IRT model applied to the national sample data already collected and analyzed by National Institute for Assessment and Evaluation of the Educational and Instructional system, INVALSI) with the actual answers to the modified ones. They then performed a qualitative analysis of some cases.

From this review of existing research methodologies, we see that there is not a standard approach to how measuring quantitative evidence for the impact of variation in the formulation of a task, nor a shared way (in a mixed-method perspective) for connecting quantitative evidence and qualitative research. The purpose of this paper is to begin to fill this gap.

### 2.3. Statistical tools

National and International large-scale surveys – such as INVALSI in Italy and OECD-PISA on a world scale - often use the Rasch Model to analyse the students' results, especially when it is needed a comparison between two different tests or the comparison between students (Barbaranelli & Natali, 2005; INVALSI, 2016; OECD, 2016). Rasch Model is a simple logistic model that belongs to the Item Response Theory (IRT) class of models. It estimates a difficulty parameter for each item of the test and an ability parameter for each student. The IRT models, and between these the Rasch Model, express the probability of giving the correct answer to an item, as a function of the item's difficulty and the ability of the student measured over the entire test.

For each item of the test, the relation between the students' ability and the probability of the correct answer is represented by a curve called *Item Characteristic Curve* (ICC). In a similar way, it is possible to use Rasch parameters to represent also the empirical data and, in particular, we can represent the trend of each possible response to an item as a function of the students' ability. These specific graphs are called *Distractor Plots* and contain many information regarding how students respond to an item.

The information gathered using the Rasch model are significant and predictive, in case of a new administration of the same test, if the students' sample size is large enough and if the statistical parameters' values' constraints (p-value and Cronbach alpha, among others) are respected. If such a prediction is possible, this information is used as a blind information about the students' expected performances with respect to an item. It is for these reasons that our statistical tool, which will be described in dephts in the next paragraph, is based on the Rasch model.

Furthermore, using the Rasch model, we are also able to apply a specific statistical procedure of test equating based on this model. In this research, we will compare the results of students obtained from two different tests that measure the same latent trait and have a common group of items- thus providing a quantitative tool for research designs like (Nesher, 1976)'s. This procedure has the task of expressing on the same scale the results of the two tests. In particular, we use a *concurrent calibration* procedure, which is considered more precise than a separated calibration (Kolen & Brennan, 2013) and allows to estimate a difficulty parameter for each item and an ability parameter for each student, considering the results of both of the tests at the same time. This kind of procedure used to link two different tests is often applied to compare results of students over time and it is called *anchoring technique*: two tests are administered to different groups of students, but the two tests contain a set of common items used to make an anchorage between the results.

### 2.4. Statement of the problem

When looking at the literature review on the impact of variations of formulation on the performance of students, we may observe that the different methodologies and results show that this impact exists and it is relevant, but there is no way for "measuring" this impact, and *a fortiori* for comparing evidences arising from different studies. Furthermore, in general it is difficult to analyse this impact in specific subgroups of students, whilst this would be important in the perspective of equity in education (for instance, if one is interested in measuring the impact of linguistic variations in mixed-languages situations).

Indeed, the effect of variations on the students' performances is not easy to investigate because the optimal situation to study is impossible to achieve. A student involved in the research should answer indeed two very similar questions, and should "forget" to have faced the first question while answering the second. Inevitably, the first task would influence the second, or the change should be so heavy to transform deeply the nature of the question itself. While a qualitative interactive *case study* of the evolution of the meaning of the text and of the *question intent* in groups or classroom discussions could suggest *a posteriori* interpretations of students' difficulties, we want to explore directly the questions: i. how much a specific variation influences the students' answers during the test? ii. What would the same student do if he should answer independently the original and the modified question? iii. What variations cause significantly different behaviours in "real-time" between two populations of students?

Hence, our research problem is to design a quantitative methodology, which integrates the existing research approaches, in order to address the two points of measuring and differentiating the impact on the students' performances due to a variation in the formulation of a problem.

In particular our research questions are:

1) It is possible to design a quantitative tool based on the Rasch model and on anchoring techniques that allows measuring the impact of a variation on the performance of the population?

A variation may cause significant changes in the answers of a population or in the performances of a particular group of students.

2) In particular, can such a tool point out from data analysis the relevance of a variation on specific sub-groups of the population?

## 3. A tool for measuring the impact of variations on students' performances

In this paper, we explain how far a quantitative approach similar to the one proposed by Branchetti and Viale (2015), integrated with suitable anchoring techniques and extended to a wider set of variations, can highlight interesting evidences which were not observed in the previous researches that we mentioned. Moreover, we show how this approach allows us to measure the impact of a variation, in relation to a well-defined scale. We present in this paragraph our methodology, based on a test linking, equating techniques and the use of the Rasch model.

### 3.1. Design of the tool

The procedure that we propose and validate is the following one. We start with a core-test (CT) composed by *n* items that measure a latent trait. The core-test, in our case, must give a statistically consistent (with respect to the statistical parameters) and mathematically significant measurement of students' ability. This core test is then considered as part of a whole test (T) composed by *m* items assessing the same latent trait. Let us denote by $A_1, A_2, \ldots, A_{m-n}$ the *m-n* items of the test (T) that do not belong to the core-test. Each one of the items $A_1, A_2, \ldots, A_{m-n}$ is then modified, by performing on it a single, well-individuated variation, giving a new set of items $A'_1, A'_2, \ldots, A'_{m-n}$. A new test T' composed by the fixed *n* items of the core test CT and the varied *m-n* items is then assembled.

We select a sample of classes, and in each class, we administer the original test T to half of the students (randomly chosen) and the varied test T' to the other half. Let us denote with P1 the population to which T is administered, and P2 the population to which T' is administered. P is then the union of P1 and P2.

The first analysis concerns only the common items (CT) of the tests T and T'. We apply the Rasch model to the CT on P, on P1 and on P2; we check the behavior of other specific statistical values, as for example the alpha-Cronbach index, for the CT on P, on P1 and on P2, in order to measure the internal consistency of the CT and its statistical validity. In this way, we have the first information about the comparability of the two sample of students who have answered to the two tests, and on the fact that adding items to the CT does not affect significantly the latent trait measured by the CT itself.

Furthermore, if the test CT has been previously administered to a particular population (for instance, to a statistical sample of a school population), it is possible to get statistical data which allow to establish comparisons between the case studied and a benchmark population. Once we have compared the results of the students on the common part of the tests, we pursue with the analysis of the other items that occur in different forms in the two tests.

The second step is performing the same analysis (Rasch model and standard statistical tools) on the test T on P1, and T' on P2, in order to measure the internal consistency of T and T' and their statistical validity.

In the third step, we reconsider the results of Rasch analysis considering only the *m-n* common items of the core-test. The ability of each student is then measured using the same items that constitute the CT, independently to the test administered. The probability that a student of a given ability level *p* (measured on the Rasch scale based on CT) answers correctly to the item Aj can be computed, and the relationship between the ability of the student and the probability of his different choices can be visualized through a graph (the above-mentioned distractor plot). We underline that, due to the characteristics of the Rasch model, this relationship includes also the information on how the choices of the students with a given ability on the test T are distributed, when answering to the item Aj (correct answer, missing, and choice of a particular distractor….).

In more detail, at this stage, we are interested in studying the different impact of the two formulations of an item ($A_1$ vs $A'_1$, $A_2$ vs $A'_2$, and so on) on the two groups P1 and P2 of students and, for this purpose, we use the Rasch ability of the students measured on the core-test. In particular, we represent how the students answers to the different versions as function of their ability; in other words, we make distractor plots of the two versions of an item ($A_i$ and $A'_i$), plotting the empirical data as functions of the ability parameter calculated on common part of the test (CT). In this way, it is possible to observe the trend of each possible answer in the two versions of an item and compare them, analysing the different behavior of the students. It is also possible to observe if this variation has a particular impact on a specific ability level and, using deeper analysis always based on distractor plots, it is possible to point out if the variation has a greater impact on a subgroup of the population (for example male or female).

Moreover, we analyse the impact of the variations also using test equating to confirm the results obtained using our procedure. We decide to use the concurrent calibration applied to the results obtained by the new administration of the two tests T and T' and this allows us to express all the parameters estimated (student's ability and item's difficulty) on the same scale. In particular, we can compare the difficulty parameters of the items $A_1, A_2, \ldots, A_m$, respectively with the parameters of the items $A'_1, A'_2, \ldots, A'_m$ and we can observe if these differences are statistically significant.

### 3.2. Output of the tool

For each pair of varied items ($A_1$ vs $A'_1$, $A_2$ vs $A'_2$, and so on) the procedure will give:

1) A (non-anchored) percentage of correct answers, of choices of distractors and of missing answers;
2) An index of difficulty for each version, placed on a common scale, anchored by the CT;
3) A distractor plot for each version, where on the x-axis the *same* ability is reported.

Roughly speaking, our tool will measure how two formulations differ as final difficulty, and how the variation of the formulation affects the performance, as a function of the ability of the student.

### 3.3. Utilization criteria

This approach needs specific controls and checks after the testing, and only after this checking one can use a quantitative tool like ours, based on statistical indicators.

UC1) First, it is necessary to verify the internal consistency of the varied tests: the Cronbach alphas of CT on P, P1 and P2 must have acceptable values. The same must happen for the Cronbach alphas of T on P1 and of T' in P2.

UC2) Second, it is necessary to verify that the three tests (CT, T and T') are related to the same latent trait- in our case, the mathematical ability. In order to verify this, it is needed a comparison of the results of the two tests in terms of distribution of students and distribution of the items, in relation to Rasch parameters. This is of course very delicate

with a purely statistical approach; in most empirical research situations, it is helpful and easier to apply a qualitative analysis observing Wright maps. We first compare the Wright maps of the core test CT when administered to P1 (as a part of T), to P2 (as a part of T'), and to P, then we do the same thing comparing the distributions of the items of CT in the Wright maps of the whole tests T and T'. If the variations have not influenced too much the core test and the latent trait measured, the distribution of the core test items must be similar in the two maps.

UC3) Third, at items' level, we must highlight if each variation has worked well and, at the same time, it is important to see if the varied items maintained good psychometrical features in relation to the core test and in relation to the whole test. To confirm that, we use the Rasch model and specific indexes of the classical test theory to analyse fit and discrimination of each item, which must fulfil the standard validation criteria.

## 4. The validation plan and its results

Our Validation Plan consists in

- Starting with a test T for which solid data are already available, coming from a large-scale assessment;
- Individuating (via a qualitative analysis of the contents and a statistical analysis of the parameters) among the items of T a core test CT, which can be assumed as a good test for measuring the mathematical ability;
- Considering variation of items, testing contents on which there are important threads of research in mathematics education, thus obtaining the test T'
- Testing the tool on a large population, comparable as characteristics to the population of the large scale assessment;
- Verifying the utilization criteria (3.3) and the comparability of the results with the results of the large-scale assessment;
- Verifying the statistical coherence and the didactic relevance of the data and the related evidences;
- Comparing these quantitative evidences with the results of the previous researches.

We underline that the cases that we will analyse are considered as a way for validating our methodology, and not as research targets; we are not trying here to interpret the didactic phenomena. We will show how our tool provide data that can be used in educational research

### 4.1. Design and administration of the tests

#### 4.1.1. The starting test

Our test T consisted of the INVALSI test administered to 590.728 Italian students of grade 6, during May 2013. We decided to use an INVALSI test as basis for our validation plan because in this way we started from a test previously administered to a large population of student and analysed in details by the INVALSI statistical team. Thereby we were sure to start from a test with good values concerning statistical reliability and coherence and that gives a statistically consistent measure of a latent trait that we can identify with mathematical ability. Furthermore, we had also the possibility to compare our results with the results of the national survey and to check if our sample is comparable to the national INVALSI sample. The INVALSI test was composed of $m = 48$ items and the statistical analysis were performed on a representative sample of 27.504 students. 1.528 of them were a representative sample of the students from the Italian region of Emilia-Romagna.

#### 4.1.2. The variations

We chose $n = 7$ items and we changed them along different directions, as described in the theoretical framework (by operating on the lexicon, the syntax, the use of figures, the registers of representation and so on). Five items are in the domain of numbers (arithmetic and estimations) and two are with geometrical content (Euclidean and analytical geometry).

Two arithmetic items are without context. In the first case, a multiplication between natural numbers without result is presented, and students are asked to choose the quantity of digits of the result of the operation among four options. The original item is a *cloze open* question and the variation concerned the typology of item: the varied item is a *multiple-choice* one.

In the second item, students are asked to choose among 4 alternatives in *multiple-choice* item, the right estimation of the result of a multiplication between rational numbers in the decimal representation. The variation here concerns the kind of numbers and their magnitude: the numbers of the varied item are obtained by multiplying the numbers by 100 (hence the numbers become natural numbers and the kind of number changes). This is a known problem in mathematics education (see, f.i., De Corte et al., 1988 and his subsequent studies). As pointed out in §2.2 the methodology used in this study is quantitative, and consist in the administration of 24 one-step problems, each proposed in two forms. The students involved were 116 and all of them solved the 24-items test twice: once in a choice-of-operation form and once in a free-response form.

Two other items in the domain of numbers are arithmetic *multiple-choice word problems* in which the data are natural numbers and the question refers to a context described verbally. In both the cases, the variation acts on the syntax of the sentences, as in the cases studied by Branchetti and Viale (2015). The fifth variation concerns the editing of the item (following Laborde, 1995).

All the original versions of the geometrical items are *word problems* with a mixed text, i.e. a text characterized by an integration between data in the verbal and in the graphic form. Two aspects of geometry are explored: a) representation of points by means of coordinates in a Cartesian plan; b) measure of lengths and areas.

In the first case, in the verbal text a path from a point (whose coordinated are reported in the text) to another is described, referring to a graphic representation. The students are asked to write the coordinates of the second point. In the varied version, the graphics representation is the same, but the coordinates of the final one are given and the students are asked to identify the coordinates of the starting point. Hence, we have here a substantial variation of the relationship between the stimulus and the task: passing from an operation to its inverse.

The second item is a problem concerning area and perimeter of composed polygons with a question in the text and the situation represented graphically. The variation concerns the first level of representation of the situation (Frank et al., 2007), which is transformed in a pure verbal one by removing the graphic representation.

#### 4.1.3. The population

We administered the new test T' and the original test T to 777 students of the same age from the same region (Emilia Romagna); they had not participated to the national assessment session in which the test T was used, in 2013. In particular, in each of the 40 classes involved in the trial, half of the students of each class (randomly chosen) answered to the new test T' and the rest of the students responded to the original test T. The validation plan is based on the analysis of the 777 tests, including 380 original test T and 397 varied test T'.

### 4.2. Data analysis - UC and comparison with the INVALSI large-scale results

In this section, we verify that our trial satisfies the utilization criteria, and at the same time we verify that the quantitative results that we obtained are coherent (along the three lines of the UC) with the results of the large-scale assessment from which T is derived - hence giving more strength to our validation.

**Table 1**
Comparison between sample-size and α-Cronbach values in the national INVALSI survey and in our trial.

| | Sample size | α-Cronbach of the whole test (T) | α-Cronbach of the core test (CT) |
|---|---|---|---|
| National INVALSI survey (test T) | 27.504 students | 0.86 | 0.85 |
| Regional restriction of the national INVALSI survey (test T) | 1.528 students | 0.86 | 0.84 |
| Populations P (only core test CT) | 777 students | Not applicable | 0.85 |
| Population P1 (test T) | 380 students | 0.86 | 0.84 |
| Population P2 (test T') | 397 students | Not applicable | 0.86 |

#### 4.2.1. UC1 and comparison with the large-scale INVALSI assessment

First, we compared the global results of our tests and the results of the national and regional original INVALSI test (Mendeley, 2018). The following table reports alpha-Cronbach of the entire original test T obtained in the national survey (considering both the national population and regional restriction of the population) compared with the alpha-Cronbach obtained in our sample, obviously considering only population P1 who responded to the original test T. Than we give also a comparison of the internal consistency of the Core Test in each of the populations taken into account (national population, regional population, P, P1 and P2).

As reported in Table 1, alpha-Cronbach values are acceptable in each test (Cronbach, 1951; INVALSI, 2016), and the internal consistency is verified both considering the whole tests (T) and analysing only the core test (CT) for each population.

#### 4.2.2. UC2 and comparison with the large-scale INVALSI assessment

Moreover, as explained in the validation criteria, we observed the distribution of the 41 items of the CT as function of the Rasch difficulty, thus obtaining that it is similar in the two tests of our trial and, even if there are minor local differences, this distribution is globally similar to that of the national and regional survey (Table 2).

Also the comparison between the Wright maps obtained analysing tests (original and varied) as a whole, allows to verify that the distributions of students and CT-items in relation to the Rasch parameter is similar and then that the presence of different items did not affect the latent trait defined by the CT-items (Table 3).

#### 4.2.3. UC3 and comparison with the statistical indexes of the original items

At last, we consider the statistical parameters of each one of the 7 items varied for the trial and we verify that their statistical parameters are still acceptable (Table 4).

In the previous table, we observe that items' parameters obtained in our trial are similar, for population P1, to those calculated on the original INVALSI test. Generally, the parameters are acceptable also in our trial and, in cases of minor anomalies (such as the too low weighted value for D18), they were already present in the National Survey.

### 4.3. Data analysis – outputs and qualitative analysis

We present here the outputs and the analysis, showing how our methodology, applied to these situations that have been studied by several authors with different approaches, provides data that integrate with what is known and can be used as a support for the interpretation of the phenomena. We recall that the outputs of our tool are the percentages of correct answers, the indexes of difficulty of the variations, anchored by means of a concurrent calibration technique, and the distractor plots relating the probability of answering correctly (to one or to the other variation) in function of the ability, measured on the same scale (Mendeley, 2018).

The complete result of this calibration in shown in Table 5, where the second column reports the calibrated index of difficulty, and the third one the standard error. This gives a measurement of the impact of the variation, which in most cases is significant. Using the values in the table below, we can calculate the coefficient $z\left(\frac{\alpha}{2}\right)$ for which the confidence intervals for the difficulty parameters of each variation, as measured by our experiment, are separated, and hence we can calculate the corresponding $\Phi(z)$ for each interval.

The values on the right in Table 6 guarantee the significance of the difference of difficulty of the questions (original and with variation) for all the questions but one analysed in our experiment.

The question D27 does not exhibit a significant impact. This case will be discussed later and compared with D16 (D16 and D27 involves related variations, both concerning linguistic factors).

#### 4.3.1. Examples of analysis 1: a number kind and size variation (item D22)

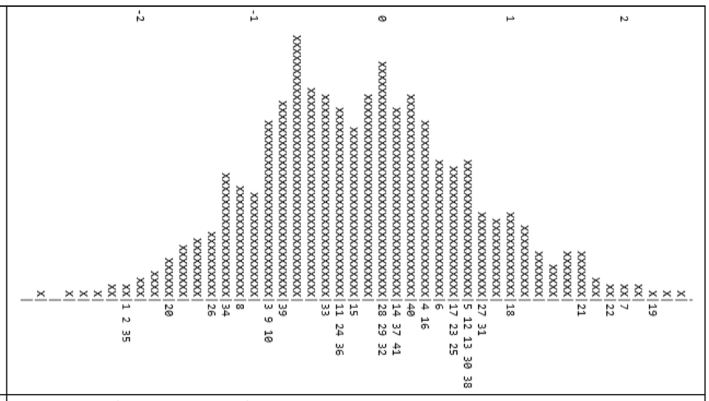
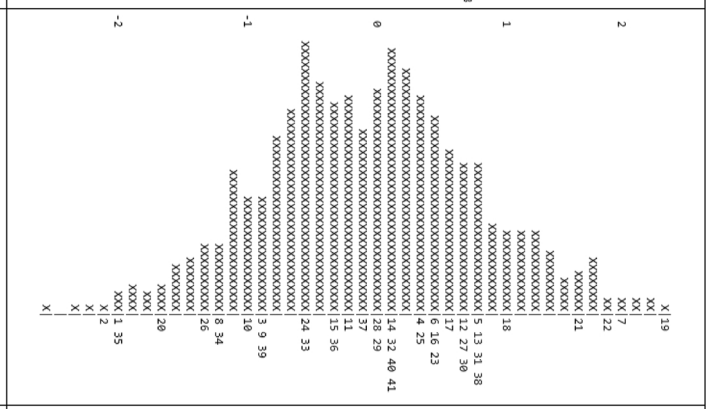
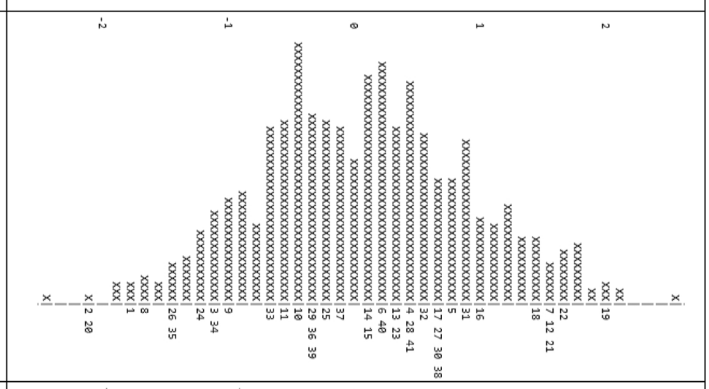
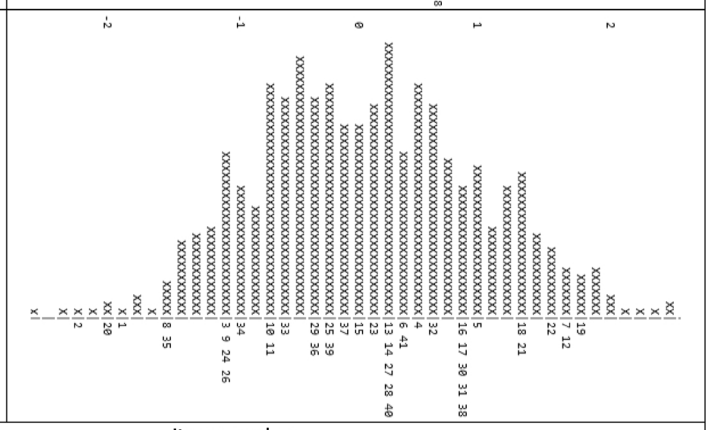
In this case (Fig. 1), the variation is numerical: changing the numbers magnitude, in this case, change the kind of numbers (from decimal to natural). We compare our variation with De Corte et al. (1988), who investigated, as we described in the background section, different students' answers when the type of numbers changed in multiplication problem-solving. It is important to notice that all the choices in the varied form are analogous to the ones in the original item: simply, each factor is multiplied by 100 and the options are multiplied by 100,000. The type and the size of the number involved is hence changed.

The *question intent* of the varied item is the same and concerns the estimate of operations results: the students are asked to choose among four options which number is closer to the result of the multiplication. As a term of comparison, we chose the paper by De Corte et al. (1988) that is mentioned as a relevant reference about arithmetical problems in the review by Daroczy et al. (2015). Even if, from a pure mathematical point of view, the kind of numbers involved shouldn't change the nature of the task we may expect, De Corte et al. (1988) showed that, in the sample they examined, there were some relevant differences. According to the authors, who studied deeply in particular students' strategies and answers when the multiplier was integer, decimal greater than 1 or smaller than 1, the percentage of students' right answers are in general more in the first case and decrease a bit moving to the second case, a lot in the third. Also, the author stressed that students who have problems in choosing the correct formal arithmetical operation, among a list of six, solving problems in which the multiplier was a decimal bigger than 1, sometimes showed to be able to estimate or, anyway, to calculate the correct result when asked to answer freely to the question on their own. The authors, relying on Fischbein, Deri, Nello, and Marino (1985), interpreted this recurrent result as a problem with the intuitive model of multiplication, that is emphasized by the contexts of the *word problems* that could "resonate", more or less, with students' models.

We conjectured that similar differences, carrying out an *a priori* analysis of the Italian curriculum, may have been observed also in our context, because of the habit to approximate with different procedures the result of a multiplication when numbers are decimal or integer. Standing on their results, we expected the performances to become better in varied case, in which the numbers are no more decimal greater than 1, but are integers.

With our methodology we are not able - and this is not the goal of this paper – to state what is the strategy, but we just want to compare our results with a solid result obtained with a different methodology in order to validate our tool and to show the additional information that it can provide about the different impact on different categories of students. The change may be ascribed indeed to differences in the teaching practices in which operations with decimal and natural numbers are used at school but, if it is the only reason why the students are driven to choose one distractor or the correct answer, the change should

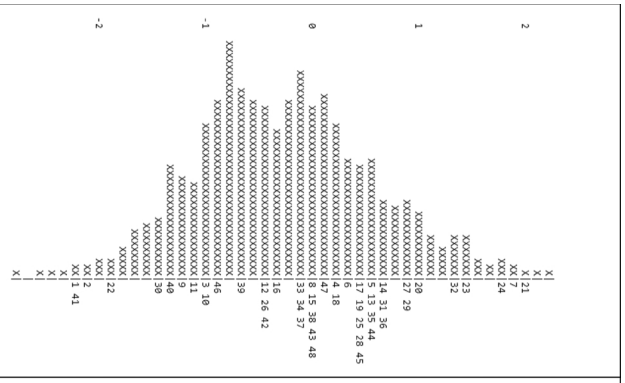
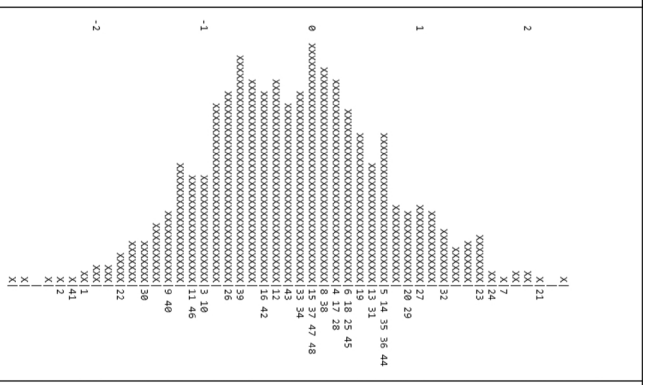
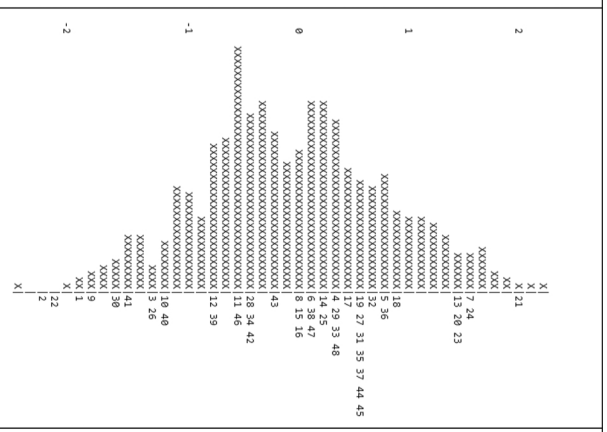
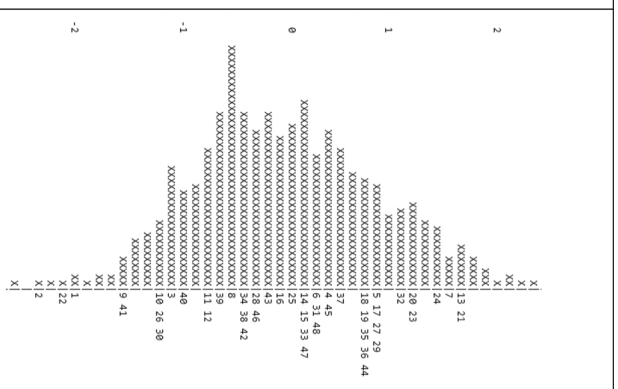
**Table 2**
Comparison between Wright map of the Core Test in the national INVALSI survey and in our trial.

| | |
|---|---|
| Wright map of the Core Test - national INVALSI survey |  |
| Wright map of the Core Test - Regional restriction of the national INVALSI survey |  |
| Wright map of the Core Test in the Original version of the test (population P1) |  |
| Wright map of the Core Test in the Varied version of the test (population P2) |  |

**Item names key:**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1_a | D1_b | D1_c1 | D1_c2 | D2_a | D2_b | D2_c | D4 | D5 | D6_a | D6_b | D6_c | D7_a | D7_b | D8_a | D9_a | D9_b | D9_c | D10_a | D10_b | D10_c |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | |
| D11 | D12 | D13 | D15 | D17_a | D17_b | D19 | D20_a | D20_b | D21_a | D21_b | D23 | D24 | D25_a | D25_b | D26_a | D26_b | D28 | D29 | D30 | |

**Table 3**
Comparison between Wright map of the Entire Test in the national INVALSI survey and in our trial.

| | |
|---|---|
| Wright map of the entire test T - national INVALSI survey |  |
| Wright map of the entire test T - Regional restriction of the national INVALSI survey |  |
| Wright map of the entire test T – new administration on population P1 |  |
| Wright map of the entire test T1' – new administration on population P2 |  |

**Item names key (items coloured in dark grey are those different in the two version of the tests):**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1_a | D1_b | D1_c1 | D1_c2 | D2_a | D2_b | D2_c | D3 | D4 | D5 | D6_a | D6_b | D6_c | D7_a | D7_b | D8_a | D8_b | D9_a | D9_b | D9_c | D10_a | D10_b | D10_c | D11 |
| **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **39** | **40** | **41** | **42** | **43** | **44** | **45** | **46** | **47** | **48** |
| D12 | D13 | D14 | D15 | D16 | D17_a | D17_b | D18 | D19 | D20_a | D20_b | D21_a | D21_b | D22 | D23 | D24 | D25_a | D25_b | D26_a | D26_b | D27 | D28 | D29 | D30 |

**Table 4**
Comparison between statistical parameters of the varied items in the national INVALSI survey and in our trial.

| | | National INVALSI survey | | Population P1 | | Population P2 | |
|---|---|---|---|---|---|---|---|
| | | Weighted | Discrimination | Weighted | Discrimination | Weighted | Discrimination |
| ITEM | D3 | 0.99 | 0.40 | 1.04 | 0.33 | 1.01 | 0.35 |
| | D8b | 1.07 | 0.31 | 1.04 | 0.30 | 1.01 | 0.39 |
| | D14 | 1.06 | 0.29 | 1.02 | 0.35 | 1.04 | 0.35 |
| | D16 | 1.09 | 0.25 | 1.05 | 0.31 | 1.12 | 0.25 |
| | D18 | 0.87 | 0.52 | 0.87 | 0.56 | 0.80 | 0.60 |
| | D22 | 1.12 | 0.25 | 1.18 | 0.15 | 1.15 | 0.22 |
| | D27 | 1.00 | 0.38 | 0.95 | 0.40 | 1.10 | 0.28 |

**Table 5**
Results of calibration procedure – difficulty of each item after test equating.

| ITEM | DELTA | ST. ERR. | ITEM | DELTA | ST. ERR. | ITEM | DELTA | ST. ERR. |
|---|---|---|---|---|---|---|---|---|
| **D1a** | -2,06 | 0,11 | **D9a** | 0,82 | 0,08 | **D20a** | -0,5 | 0,08 |
| **D1b** | -2,4 | 0,12 | **D9b** | 0,63 | 0,08 | **D20b** | 0,65 | 0,08 |
| **D1c1** | -1,25 | 0,09 | **D9c** | 1,32 | 0,09 | **D21a** | 0,75 | 0,08 |
| **D1c2** | 0,38 | 0,08 | **D10a** | 1,84 | 0,1 | **D21b** | 0,5 | 0,08 |
| **D2a** | 0,77 | 0,08 | **D10b** | -2,26 | 0,11 | **D22o** | 0,1 | 0,11 |
| **D2b** | 0,19 | 0,08 | **D10c** | 1,33 | 0,09 | **D22v** | -0,51 | 0,11 |
| **D2c** | 1,56 | 0,1 | **D11** | 1,49 | 0,09 | **D23** | -0,76 | 0,08 |
| **D3o** | 0,03 | 0,11 | **D12** | 0,09 | 0,08 | **D24** | -1,18 | 0,09 |
| **D3v** | -0,57 | 0,11 | **D13** | -1,32 | 0,09 | **D25a** | -1,63 | 0,09 |
| **D4** | -1,79 | 0,1 | **D14o** | 0,52 | 0,11 | **D25b** | -0,47 | 0,08 |
| **D5** | -1,24 | 0,09 | **D14v** | 0,88 | 0,12 | **D26a** | -0,25 | 0,08 |
| **D6a** | -0,74 | 0,08 | **D15** | -0,38 | 0,08 | **D26b** | 0,59 | 0,08 |
| **D6b** | -0,82 | 0,08 | **D16o** | 0,31 | 0,11 | **D27o** | 0,55 | 0,12 |
| **D6c** | 1,53 | 0,1 | **D16v** | 0,8 | 0,12 | **D27v** | 0,37 | 0,11 |
| **D7a** | 0,12 | 0,08 | **D17a** | -1,47 | 0,09 | **D28** | -0,45 | 0,08 |
| **D7b** | 0,03 | 0,08 | **D17b** | 0,38 | 0,08 | **D29** | 0,06 | 0,08 |
| **D8a** | -0,1 | 0,08 | **D18o** | 0,68 | 0,12 | **D30** | 0,3 | 0,08 |
| **D8bo** | 0,47 | 0,11 | **D18v** | 1,09 | 0,12 | | | |
| **D8bv** | 0,77 | 0,12 | **D19** | 0,24 | 0,08 | | | |

**Table 6**
Values of z and $\Phi(z)$ for the impact of the variation in each question.

| Question | Z | $\Phi$(z) |
|---|---|---|
| D3 | 2.72 | 0.997 |
| D8 | 1.30 | 0.903 |
| D14 | 1.57 | 0.942 |
| D16 | 2.13 | 0.983 |
| D18 | 1.71 | 0.956 |
| D22 | 2.77 | 0.997 |
| D27 | 0.78 | 0.782 |

influence all the students in the same way. On the contrary, one might observe that the students, once overpassed an intermediate Rasch score, are not conveyed by this kind of variation. We can investigate this additional issue with our methodology.

The results by De Corte et al. (1988) are possible to compare with our samples because there are important similarities and compatibilities:

1 We asked the students to answer in a written form to the questions, like De Corte et al. (1988).
2 We propose two formulations in which decimal numbers bigger than 1 and integers are used as multipliers; these differences might impact lower than the other (decimal multiplier smaller than 1) and our methodology might help to make clearer how they impact on a whole population and on specific subgroups.
3 We didn't propose *word problems* with a context but we just asked to estimate the result of a multiplication; this way we avoid also the interference of intuitive models linked to the contexts.
4 We ask the students to choose among different results but not to choose the formal arithmetic operation, so the students are free to choose the result in the way he/she prefers and the eventual differences in the performances between the answer to the original and the varied items are not due to this specific difficulty; this way the only variable (size of multiplier) becomes the type of number used as a multiplier and not the response mode, harder to investigate and so more critical to use in a validation plan.

The main differences are the following:

1 the students in De Corte et al. (1988) were the same and answered

| | |
|---|---|
| Original form | **D22. Which of the following integer numbers is closer to the result of this multiplication?**<br>**4,82 x 9,95**<br>**A. 36**<br>**B. 42**<br>**C. 48**<br>**D. 50** |
| Varied form | **D22. Which of the following integer numbers is closer to the result of this multiplication?**<br>**482 x 995**<br>**A. 360.000**<br>**B. 420.000**<br>**C. 480.000**<br>**D. 500.000** |

**Fig. 1.** D22 – item in the original form (test T) and in the varied form (test T').

**Table 7**
Answers percentages for item D22 (original form and varied form). Correct answer: C.

| | Original item | Varied item |
|---|---|---|
| A | 15% | 11% |
| B | 21% | 13% |
| **C** | **46%** | **59%** |
| D | 11% | 12% |
| MISSING | 6% | 6% |

twice to the test with different response modes a week later; we compared different students with the same feature and every student answered once.

2 their methodology was both quantitative and qualitative, while our is only quantitative.

For what we are looking for, the differences between the methodologies are not such to invalidate the comparison, since we would have considered the same students answering twice as "two equivalent students" according to our classification. We decided to analyse this kind of variation in order to validate our measurement methodology, since it allowed to study a situation where a change was expectable with high probability and the result is a solid finding that was also partially interpreted by the authors. Our methodology, according to our criteria, should have at least confirmed this result to be considered valid from an educational point of view. The data analysis, performed as we have explained before, demonstrates that with this variation the item becomes much easier, confirming the hypothesis and the results by De Corte et al. (1988). Indeed, the percentage of correct answers in the original item is 46% and in the varied one increases to 59% (Table 7).

As we can see in the table before, the variation has mostly influenced the choice of distractor B. Whilst the distractors A and D increase or decrease only of some percentage points, answer B loses about 8% in the varied item. The variation did not influence the percentage of students who did not answer to the item and this may mean that, despite the different difficulty of the two versions, almost all the students are confident enough to try to answer.

The *concurrent calibration technique* applied to both the tests simultaneously allows us to estimate the difficulty parameters of all the items (including the two versions of the 7 items modified) and to consider them on the same scale. The comparison between difficulty parameters estimated in this way gives us an additional proof that the variation in this case had made the item easier. In fact, after using the anchoring procedure, the difficulty parameter of the original item is 010 and the difficulty of the varied item is -0,51, both with a standard error of 0,11, hence the difference is statistically significant.

At this point of our analysis, it is interesting to use Rasch analysis and, in particular, distractor plots to investigate if the differences identified before are distributed in the same way on all the students or if these changes have influenced in particular students with a certain ability level. Using the Rasch Model to analyse the core test (CT) composed by 41 common items, we estimate an ability parameter for each of the 777 students. The distractor plots below (Fig. 2) represent the empirical data of the two versions of the item D22 as function of the ability of the students evaluated before using the core test.

The behaviors of the correct answer and of the wrong choices is different in the two items. In particular, the trend of the correct answer (green) is more regularly increasing in the varied item. In both of the two forms, we observe that this question is not very discriminant (in the sense of ITR). In other words, this item does not distinguish well between students with high ability levels and students with lower ones. This information is also confirmed by the analysis of the whole tests: the discrimination parameter of this item is 0.15 for what concerns the original test and 0.22 in the varied test (hence slightly better). Also in the national survey this item has a discrimination below the threshold (0.20), but it is interesting to notice that the variation improved the statistical properties of the item. Indeed, the discrimination is higher in the varied form of the item and also comparing the weighted index of the two items, the varied one gives better results.

In this paper we are interested in the analysis of the comparison between the two items and, in particular, useful information emerge from the comparison between distractor trends. For example, we can focus on the trend of the distractor B (blue), which is the one with the most relevant variation in the percentage seen before: by comparing the two plots, we can see that distractor B was often chosen by students with high ability level in the original item, while the variation made it much less attractive for those students. In respect to De Corte et al. (1988) we have more information about the distribution of the answers, since on different categories of students the variation had different impacts, but also about the kind of students (in this case in terms of ability level) that are more influenced by the variation and the kind of answers that are not considered in one case and in the other. We can conjecture that students with high levels of ability in the first case multiplied the integers and then estimated just the decimal rest smaller than 1, making mistakes in this last step maybe because of the intuitive model of multiplication (a multiplication between two numbers smaller than 1 should be "small"), while in the second case this was not used as a strategy because the students worked with integer numbers.

Furthermore, the analysis of this item is very interesting also because the variation has a very different impact on males and females. In the table below the percentages of each answer for male and female and for both the two versions of the item are listed. In the original item with the decimal numbers, correct answers are the 44% for males and 50%

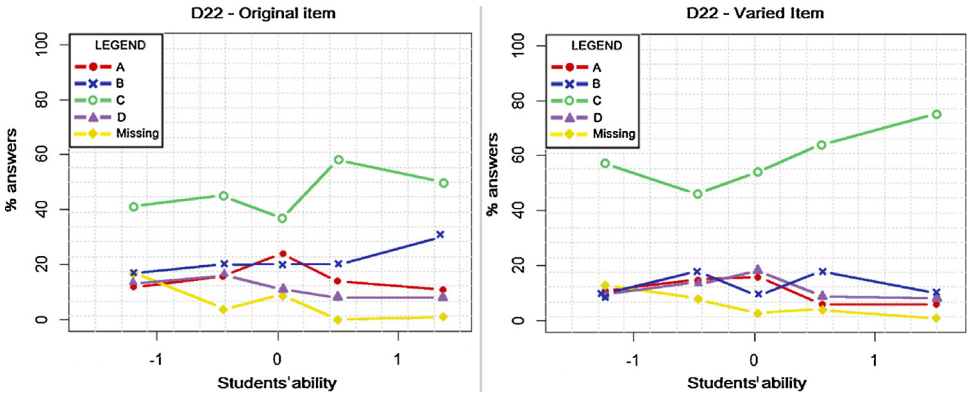



**Fig. 2.** D22 – Distractor plots: original item (D22_o) and varied item (D22_v).

**Table 8**
Answers percentages for item D22_o and D22_v for male and female. Correct answer is C.

| | MALE | | FEMALE | |
|---|---|---|---|---|
| | original item | varied item | original item | varied item |
| A | 16% | 10% | 12% | 11% |
| B | 23% | 12% | 20% | 13% |
| **C** | **44%** | **62%** | **50%** | **56%** |
| D | 11% | 11% | 10% | 13% |
| MISSING | 5% | 4% | 8% | 7% |

for females. The variation has a huge impact on males' performances (Table 8) and, answering to the varied item, 18% more of the students choose the correct answer. Instead, if we observe female percentages, we can see that the correct answer takes only the 6% more in the varied version.

This evidence is a completely new result, never explored in the literature concerning mathematical problem solving and also in *gender gap* researches. Our methodology, just varying the variable used to create categories within the groups, allowed us thus to make new phenomena emerge that could become new research issues to explore with suitable mixed methodologies. We provide just a conjecture to interpret the quantitative result. This might mean that males and females apply different strategies to solve this item, or that they differently adapt these strategies to a different situation: the strategies used by females improve their performance less than for males even if the numbers are changed. As a further issue, it might be interesting to investigate with our technique more deeply quantitative evidences about gender gap in mathematics task (see Bolondi, Cascella & GIberti, 2017).

#### *4.3.2. Example of analysis 2: linguistic variations (items D16 and D27)*

In both cases, the items are arithmetic *multiple-choices word problem* and the variation concerns a linguistic variation: the varied item has a more complicated periods syntax.

In the first item, the variation concerns the syntactic level and is a complication of the period's structure. A linguistic analysis of the original text shows that it is composed by two periods and that, while the first has just a principal sentence, the second has a principal sentence including the question and a relative subordinate sentence. In the modified item, the periods are joint. The principal sentence contains the question and there are subordinate sentences, depending on it, that are a relative one and a conditional one, which depends on a declarative subordinate (see Fig. 6). We reproduce here, for obvious reasons, the Italian text (Figs. 3 and 4 ).

The second item too was modified at the syntactic level, but in a different way (Fig. 5). The variation concerns again the syntactic level and is a reduction of the number of sentences and an inversion of order between the question and the data. This variation is thus syntactic but affects also the logic of resolution, since the data are once before (original), once after (varied) the question.

The original text is composed by three periods. The first is composed by a principal sentence and three relative subordinates at the same level. The other two periods are composed by principal sentences, and the second contains the question. The varied item has one period less (just two) and the last is a complex period with a principal sentence-which contains the question- and three relative sentences, containing the data.

The variations introduced morpho-syntactic structure that are typical of *word problems* in Italian textbooks (Bolondi & Viale, 2017), making it *a priori* more "familiar" to students.

In the first case (D16): a) we moved the question, that was at the end, to the beginning of the text; b) we reduced the quantity of sentences and we used a subordinate sentence introduced by a gerund in order to introduce two relevant data ("knowing that…") in the period containing the question.

In the second case (D27): a) we moved the question, that was at the end, to the beginning of the last sentence; b) we used a subordinate sentence introduced by a gerund to introduce three relevant data ("knowing that…") in the period containing the question.

According to the description by Frank et al. (2007), such variations could have an impact on students' access to the *surface of the problem* (first level) and on students' identification of the relationships between the elements and the connections, since we changed the relationships between the main elements of the text's network (the *textbase*, second level). The students' knowledge might be involved differently in the two cases, according to the way the other two levels of representation allowed the students to recall their resources. In general, we expected, standing on this solid finding, different distributions among the different groups of students with different ability levels, while we expected similar impacts of the variation on the whole population, since the kind of variations were very similar in the two cases. Branchetti and Viale (2015) showed that, increasing progressively the difficulty of the mathematical task, the linguistic variations had more impact on students with lower levels of ability, but that in some cases students with a very good ability level in mathematics made mistakes facing items varied syntactically even if they were expected to solve them correctly in the original version, according to their level of competence. We explored the students' performances variations, through a "double comparison" between two items varied in analogous ways, looking for quantitative phenomena that, first of all, confirmed the result that linguistic variation changed the percentage and the distributions of correct answers in both the cases (confirmation of the result) and, in addiction, gave new insights about their impacts on students with different ability levels, checking the hypothesis by Branchetti and Viale (2015) about the impact of some syntactic variations on good students' performances.

Looking at the global parameters of the items, some general trends

| | |
|---|---|
| Original form | **D16. Una scatola di cioccolatini contiene 15 cioccolatini al latte e 25 cioccolatini fondenti. Con 100 cioccolatini al latte e 180 fondenti, qual è il numero massimo di scatole con la stessa composizione della precedente che si possono riempire?**<br>A. ☐ 5<br>B. ☐ 6<br>C. ☐ 7<br>D. ☐ 8 |
| Varied form | **D16. Qual è il numero massimo di scatole di cioccolatini che si possono riempire con 100 cioccolatini al latte e 180 fondenti, sapendo che ogni scatola deve contenere 15 cioccolatini al latte e 25 fondenti?**<br>A. ☐ 5<br>B. ☐ 6<br>C. ☐ 7<br>D. ☐ 8 |

**Fig. 3.** D16 – item in the original form D16_o (test T) and in the varied form D16_v (test T').

emerged. The parameter of difficulty of the items changed with the variation of the formulations in both the cases, as much as the percentage of right answers, even if in the second case the impact of the variation is not statistically significant (and this resonates with the disconfirmation of our hypothesis, as we will explain in the following). In the first case, the value of the parameter of difficulty changes from 0.31 to 0.80 and the percentage of right answers decreases from 41.58% to 33.58%. In the second case, we observe an opposite trend: even if the variation was expected to make the sentence more difficult, the value of the parameter of difficulty changes from 0.55 to 0.37 and the percentage of correct answers increases, from 36.84% to 40.55%. In Tables 9 and 10 the global results for every option in the two cases are reported.

We propose here a comparison between the distractor plots obtained categorizing the students by levels of ability, in order to deepen the data analysis (Fig. 6).

The correct answer is represented by a blue line. The yellow ones represent the missing answers, while the other represent the other options.

Comparing the graphs of the original versions, we observe first that the graphs of the formulations D16_o and D27_o are very different for students with the lowest performances, but with significant resemblances for students with medium-high performances.

The graphs corresponding to the formulations with less periods and hence a more complex syntactical structure, shows different changes in students with the same level of competence in the test: i.e., different variations affected the same category of students in different ways. Let us analyse some emergent features of the graphs.

The students with the highest performances had the 60% of right answers to both D16_o and D27_o, but had significantly different performances with the second formulation (around 40% in D16_v and around 60% in D27_v). In particular, in the first case there is an evident difference between D16_o and D16_v, while in the second case the percentage is almost the same in D27_o and D27_v for what concerns students with high performances, so the variation did not influence their performances. On the contrary, in the case of students with an average competence no significant effects have been found. The students with lower performances in the whole process had a more complex behavior. In the case of D16_o and D16_v, the impact of having a more complex syntactical formulation (D16_v) was to reduce the percentage of right answers and to make the answers distribution over the 4 options almost casual (the options have very similar percentages). In the second case (D27_o and D27_v), the percentage of right answers increases a lot with the variation.

Since the statistical parameters are good in the first case, we consider such an anomaly interesting and maybe due to relevant factors in the second case, that could be interesting to investigate from a didactical point of view. This result suggests and encourages us to investigate better this phenomenon, in order to clarify the relationships between formulation, the mathematical competence, linguistic skills and assessment. Further investigations should be carried out with qualitative methods, like case studies, interviews and focus groups analysis with students, analysis of textbooks and teaching practices concerning arithmetic *word problems*.

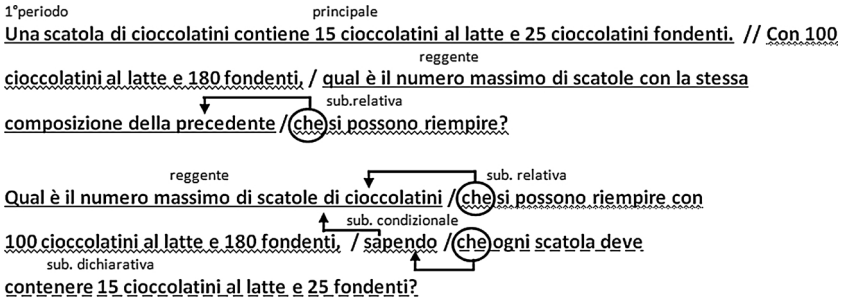



**Fig. 4.** Syntactical analysis of D16_o and D16_v.

| | |
|---|---|
| Original form | **D27. Nello zaino di Chiara ci sono il libro di scienze, che pesa mezzo chilo, il libro di matematica, che pesa 980 g, e due quaderni uguali. Libri e quaderni pesano in tutto due chilogrammi. Quanto pesa ciascun quaderno?**<br>A. ☐ 150 g<br>B. ☐ 260 g<br>C. ☐ 510 g<br>D. ☐ 520 g |
| Varied form | **D27. Nello zaino di Chiara ci sono il libro di scienze, il libro di matematica e due quaderni uguali. Quanto pesa ciascun quaderno, sapendo che il libro di scienze pesa mezzo chilo, il libro di matematica pesa 980 g, e che libri e quaderni pesano in tutto due chilogrammi?**<br>A. ☐ 150 g<br>B. ☐ 260 g<br>C. ☐ 510 g<br>D. ☐ 520 g |

**Fig. 5.** D27 – item in the original form D27_o (test T) and in the varied form D27_v (test T').

**Table 9**
Parameter of difficulty and percentages of the original and varied versions of D16. Correct answer: B.

| D16_o - Original item | | | D16_v - Varied item | | |
|---|---|---|---|---|---|
| | Count | % of tot | | Count | % of tot |
| **A** | 49 | 12.89 | **A** | 43 | 10.83 |
| **B** | **158** | **41.58** | **B** | **131** | **33.00** |
| **C** | 108 | 28.42 | **C** | 132 | 33.25 |
| **D** | 54 | 14.21 | **D** | 62 | 15.62 |
| **Missing** | 11 | 2.89 | **Missing** | 29 | 7.30 |
| Item (anchored) Delta: 0.31 | | | Item (anchored) Delta: 0.80 | | |

**Table 10**
Parameter of difficulty and percentages of the original and varied versions of D27. Correct answer: B.

| D27_o - Original item | | | D27_v - Varied item | | |
|---|---|---|---|---|---|
| | Count | % of tot | | Count | % of tot |
| **A** | 87 | 22.89 | **A** | 45 | 11.34 |
| **B** | **140** | **36.84** | **B** | **161** | **40.55** |
| **C** | 48 | 12.63 | **C** | 56 | 14.11 |
| **D** | 49 | 12.89 | **D** | 73 | 18.39 |
| **Missing** | 56 | 14.74 | **Missing** | 62 | 15.62 |
| Item (anchored) Delta: 0.55 | | | Item (anchored) Delta: 0.37 | | |

## 5. Conclusions and further issues

The analysis performed in sections 4.2 and 4.3 allows us to state that our tool, tested in a situation where many variables were controlled a priori (by means of the checked relationship between our test and our population with a large-scale assessment), gave back coherent data on the impact of different kinds of variations in the formulation of a problem. This impact has been measured and it turned out to be statistically significant in the cases of D22 and D16. The results of our measure were interpreted through a didactical *a priori* analysis based on the existing literature concerning, time after time, the particular variations. This analysis allowed to highlight phenomena already known in the literature, and to point out new aspects of these phenomena (for instance, their different relevance for different subgroups of students).

Our analysis, nevertheless, shows also that this quantitative method, in order to give useful information, needs to be integrated with qualitative analyses. In this sense, we need to understand which mixed research methodologies can give the better results and if it is possible a unique qualitative approach for different variations, or if is necessary to find specific methodologies in each case – the textual or syntactic variation could be investigate differently than a variation involving different computation strategies induced by different numbers or a variation about the graphical manipulation induced by the presence of a graphical representations. Moreover, the statistical requests for the validity of the tool are rather important and not easy to satisfy: the tool requires a large number of students in order to be effective, especially if one wants to use it for studying subgroups. Hence the variations must be deeply studied, a priori, before implementing the experimental plan, since the data collection may take a lot of time in order to have a suitable number of students involved.

In synthesis, we may state our answers to the research questions posed in §2.4.

1) Yes. Our tool exploits the specific features of Rasch model and anchoring techniques and has been validated in significant and controlled situations. Its intrinsic limit consists in the dimension of the experimental plan needed for implementing it.
2) Yes. In our situations, the tool pointed out relevant new phenomena for particular subgroups of the population, thus suggesting new research themes.

**Funding information**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Declaration of interests**

None.

**Acknowledgments**

The authors thank Matteo Viale for the careful readings of the first

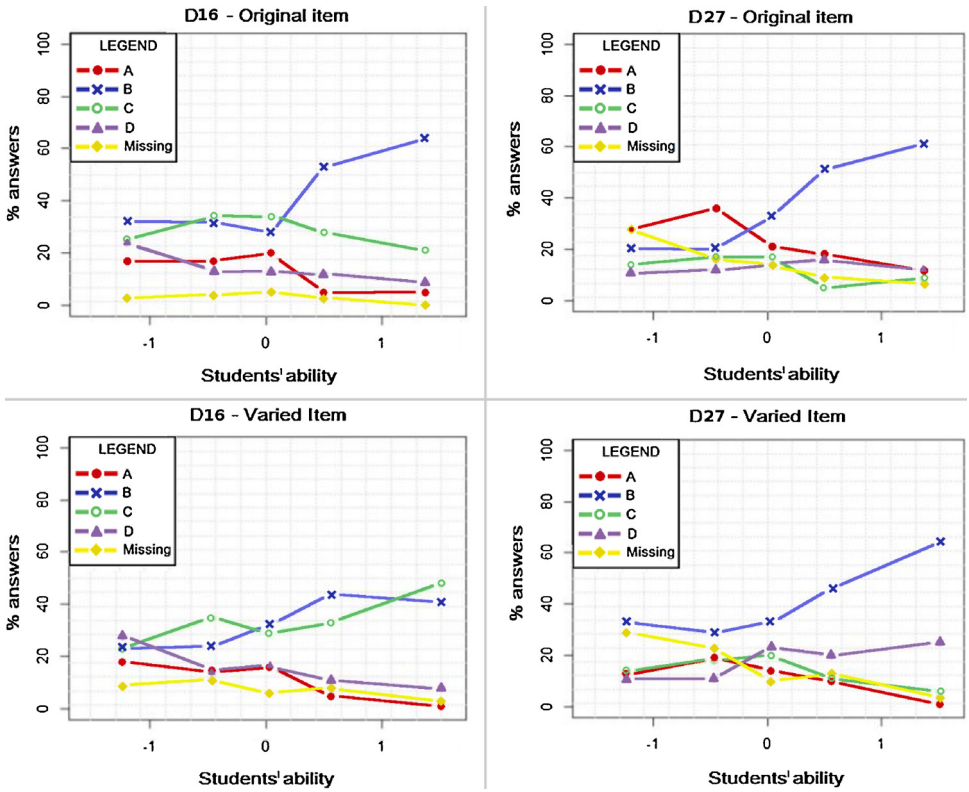



**Fig. 6.** Distractor plots of the D16 (left) and the D27 (right), original (up) and varied (down) items.

version of this paper and helpful remarks; Clelia Cascella for her assistance on the statistical aspects; Rebecca Boninsegna and Alice Lemmo who participated to the experimental part of the research; the teachers of the schools who collaborated to the data collection of our research, giving a fundamental contribution to the whole investigation. This work was supported by the Open Access Publishing Fund provided by the Free University of Bozen-Bolzano.

## References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234. http://dx.doi.org/10.1207/S15324818AME1403_2.

Bagni, G. T., & D'Amore, B. (2005). Epistemologia, sociologia, semiotica: la prospettiva socio-culturale. *la matematica e la sua didattica, 1*, 73–89.

Barbaranelli, C., & Natali, E. (2005). *I test psicologici: teorie e modelli psicometrici.* Roma: Carrocci Editore.

Bolondi, G., & Viale, M. (2017). Abilitàlinguistiche e discipline scientifiche: un'esperienza di formazione del corpoinsegnante nel Polo dell'Emilia-Romagna del progetto "I Lincei per unanuova didattica nella scuola" *Educazione linguistica eapprendimento/insegnamento delle discipline matematico-scientifiche* (pp. 173–185).

Bolondi, G., Cascella, C., & Giberti, C. (2017). Highlights on gender gap from Italian standardized assessment in mathematics. *SEMT 17 proceedings – International Symposium Elementary Maths Teaching* 81.

Branchetti, L., & Viale, M. (2015). Tra italiano e matematica: il ruolo della formulazione sintattica nella comprensione del testo matematico. In M. Ostinelli (Ed.). *La didattica dell'italiano. Problemi e prospettive* Proceedings of the conference "Quale didattica dell'italiano? Problemi e prospettive", Locarno, ottobre 2014.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405–438.

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology, 6*, 348. http://dx.doi.org/10.3389/fpsyg.2015.00348.

D'Amore, B. (2000). Lingua, matematica e didattica. *la matematica e la sua didattica, 1*, 28–47.

De Corte, E., Verschaffel, L., & Van Coillie, V. (1988). Influence of number size, problem structure, and response mode on children's solutions of multiplication word problems. *J. Math. Behav. 7*, 197–216.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77*(4), 460–470.

Duval, R. (1991). Interaction des différents niveaux de représentation dans la compréhension de textes. *Annales de Didactique et de sciences cognitives,* 136–193.

Education Committee of the EMS (2011). Solid findings mathematics education. *Newsletter of the European Mathematical Society, 81*, 46–48.

Fischbein, E., Deri, M., Nello, M., & Marino, M. (1985). The rule of implicit models in solving verbal problems in multiplication and division. *Journal of Research in Mathematics Education, 16*, 3–17.

Frank, S. L., Koppen, M., Noordman, L. G., Vonk, W., & Perfetti, C. A. (2007). *Modeling multiple levels of text representation. Higher level language processes in the brain: Inference and comprehension processes.* 133–157.

INVALSI (2016). *Rilevazione nazionale degli apprendimenti 2015-2016. Rapporto tecnico.* Retrieved July 2017 fromhttp://www.invalsi.it/invalsi/doc_evidenza/2016/002_Rapporto_tecnico_2016.pdf.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Kolen, M. J., & Brennan, R. L. (2013). *Test equating: Methods and practices.* Springer Science and Business Media.

Laborde, C. (1995). Occorre apprendere a leggere e scrivere in matematica. *la matematica e la sua didattica, 9*(2), 121–135.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21*(1), 83–90.

Mendeley (2018). Research data for validating a quantitative methodology to analyse the impact of text formulation on students learning assessment in mathematics, Mendeley Data, v2. https://doi.org/10.17632/p56btpvkrd.2.

Nesher, P. (1976). Three determinants of difficulty in verbal arithmetic problems. *Educational Studies in Mathematics, 7*(4), 369–388.

Nesher, P. (1982). Levels of description in the analysis of addition and subtraction word problems. *Addition and Subtraction: A Cognitive Perspective,* 25–38.

OECD (2016). *PISA 2015 assessment and analytical framework: Science, Reading, mathematic and financial literacy.* Paris: OECD Publishing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut.

Schleppegrell, M. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 23*(2), 139–159.

Spanos, G., Rhodes, N. C., & Dale, T. C. (1988). Linguistic features of mathematical problem solving: Insights and applications. *Linguistic and Cultural Influences on Learning Mathematics, 221*.

Thevenot, C., Devidal, M., Barrouillet, P., & Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. *The Quarterly Journal of Experimental Psychology, 60*(1), 43–56.

Thevenot, C., & Oakhill, J. (2005). The strategic use of alternative representation in arithmetic word problem solving. *Quarterly Journal of Experimental Psychology, 58*, 1311–1323.

Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology, 77*(4), 829–848.