Real tasks in Italian mathematics standardized assessment

Chiara Giberti, Simone Passarella

Real tasks in Italian mathematics standardized assessment

Chiara Giberti[1] and Simone Passarella[2]

[1]University of Bergamo, Department of Human and Social Science, Bergamo, Italy; chiara.giberti@unibg.it

[2] University of Padova, Department of Mathematics, Padova, Italy; simone.passarella@unipd.it

In this research we propose a criterion to classify real mathematics tasks in a standardized assessment. Specifically, we used this criterion to categorize the tasks of six, grade 5, standardized tests administered in Italian schools from 2009 to 2014 and we analyzed the psychometric features of these items. Results show a consistent presence of real tasks in Italian standardized assessment, indicating that one of the competences that these tests are intended to assess is the ability to develop and apply mathematical thinking to solve problems from everyday situations, as desired by the European Council and PISA. Moreover, the items classified as real are not necessarily difficult for students, but they are discriminative. The analysis at an item level of this kind of tasks could provide further information about how students face real tasks in a standardized assessment

Keywords: Standardized assessment, rasch model, real task, mathematical competence.

Introduction

In recent years, the growing attention given to national and international standardized assessment is promoting new reflections on how these tests are constructed and which information emerge from their results (Giberti & Maffia, 2020). The theoretical framework of each survey is fundamental to understand the aim following which the tasks are constructed and the mathematics that the test is going to assess. The most famous and discussed international survey is PISA, the Programme for International Student Assessment, promoted by the Organization for Economic Co-operation and Development (OECD). In PISA theoretical framework, Mathematical literacy is defined as "an individual's capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts" (OECD, 2019); then all PISA mathematics problems are set in a context, which can be personal, occupational, societal, or scientific.

This contribution reports first results of an ongoing research project whose overall aim consists in studying how tests of Italian standardized assessment (administered by the Italian National Institute for the Evaluation of Educational Systems – INVALSI - every year at grade 2, 5, 8, 10 and 13) are built and what mathematical competencies and types of reasoning are aimed to be assessed. Specifically, the aim of this paper is to analyze INVALSI tests, proposing a criterion based on Palm's framework (Palm, 2006) to evaluate the presence and quality of real tasks. Furthermore, the psychometrical features of those tasks previously identified as real will be analyzed.

Theoretical Framework

In the Italian context, the National Indications for the First Cycle of Education (MIUR, 2012), emphasize how mathematical knowledge should offer skills for perceiving, interpreting, and linking artifacts and daily-life events. Students are required to analyze situations and to translate them into mathematical terms and choose the actions to be performed to produce a solution to the problem.

These indications are reflected also in the European context. Indeed, the European Council (Recommendations of the European Parliament for lifelong learning (2018/C189/01)) recently defined as *mathematical competence* the ability to develop and apply mathematical thinking and the ability to use different representations (formulas, constructs, models, …) to solve problems starting from everyday situations. Despite the importance of including real-life experiences in mathematics classrooms, there is a strong discontinuity between in- and out-of- school mathematical experiences (Passarella, 2021). One of the possible causes of this fracture is the stereotyped nature of the problems proposed by textbooks and teachers, which, rather than serving as an interface between mathematics and reality, promote in students an exclusion of realistic considerations (Verschaffel et al., 2020). Instead, real and less stereotyped problems must be inserted in the school practice, in order to create a bridge between mathematics classroom activities and everyday-life experiences (Passarella, 2021; NCTM, 2000). Given the importance of real problems for mathematics classrooms, the following question emerges: when a mathematical task might be defined as *real*? Several studies delt with the concordance between mathematical school tasks and the corresponding out-of-school situations: Wiggings (1993) with the introduction of the term authenticity; Niss (1992) and the mathematical literacy framework of the OECD's Program for Student Assessment, PISA (OECD, 1999); Realistic Mathematics Education, that considers realistic and rich contexts as starting points for mathematization processes (Van den Heuvel-Panhuizen & Drijvers, 2014). In this contribution, we will consider the framework proposed by Palm (2006), that was developed to depict those aspects of real-life situations that should be considered important in their simulations (Figure 1).

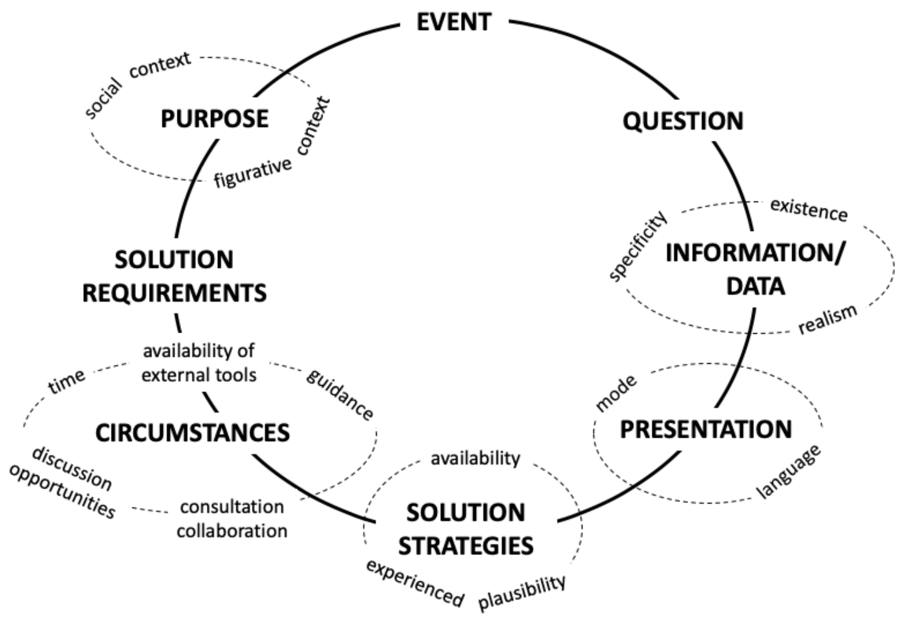


Figure 1: Palm's framework for simulations of real-life situations

A detailed description of these aspects can be found in Palm (2006). In Table 1 we describe those aspects that will be considered in the rest of the work, reporting for each of them a key question, in order to start reasoning on the concordance between mathematical school tasks and corresponding situations in real-life beyond the mathematics classroom. The framework proposed by Palm can enhance discussions amongst teachers and researchers on the characteristics of tasks that simulate real-world situations, and for the design and analysis of real tasks for mathematics classroom activities (Palm, 2006). Starting from Palm's framework, in this work we are proposing a criterion to evaluate the presence of real tasks in the Italian national tests in mathematics. Furthermore, using this criterion, we identify real task and analyze their psychometrical features to understand the behavior of the students facing these items.

Table 1: Key questions for some aspects from Palm's framework

Event	Could the event described in the school task take place in real-life?
Question	Does the questions in the school task might be posed in a corresponding real-life event?
Existence	Do the information available in the school task exist in a corresponding real-life event?
Realism	Are the values given in the school task close to values in a corresponding real-life event?
Specificity	Can specifications of the school task context be compared to a reasonably extent to the corresponding out-of-school situation?
Presentation Mode	Is the problem communicated orally or in a written form? Are the information presented in diagrams, tables, graphs, …?
Language	Does the language used in the school task not negatively affect the possibilities for students to use the same mathematics as they would have used in a corresponding real-life event?

Research Questions

The first aim of this contribution is to evaluate the presence of real tasks in the Italian national tests in mathematics (INVALSI tests), in order to investigate if INVALSI tests are intended to assess students' ability to develop and apply mathematical thinking to solve problems starting from everyday situations, as desired by the European Council and PISA. Consequently, our first research question is:

RQ1: Are there real tasks in the standardized INVALSI tests?

If the answer to the first research question is yes, considering that previous studies have highlighted that there is a strong discontinuity between in- and out-of- school mathematical experiences (Passarella, 2021), we would investigate the psychometrical features of the tasks classifies as real, in order to start understanding how students face these tasks.

Accordingly, our second research question is:

RQ2: If yes, which are the psychometrical features of these tasks? Are there some recurrent features that can be interpreted as a specific behavior of students facing a task of this type?

In this paper we will answer to this second research question considering the general psychometrical features provided by the Rasch Model, considering the whole set of items and providing an example of analysis at an item level.

Data and Methods

INVALSI tests are national standardized assessment administered every year in different school grades from primary to upper secondary school. In all grades INVALSI tests investigate students' competencies in mathematics and Italian language but, in upper grades, also an English test was included in the survey. In this work we decided to focus on grade 5 mathematics tests and compare the results in different years: from 2009 to 2014. INVALSI test are administered, for each grade, to the entire student's population. In this research our analysis is based on the probabilistic national sample drown from the entire population which is also considered for all the statistical analysis of the INVALSI Institute. For each test the sample includes approximately 30 000 students[1] and sample data are completely void of the effect of cheating because the test administration is conducted by an INVALSI inspector who guarantees the correctness and fairness of the process. All mathematics tests include different items in terms of typology (Multiple choice, Open-ended, Justification Open-ended, True/False), and mathematical content categories (Numbers, Data and Probability, Relationships and Functions, Space and Shape).

To answer to our first research question, we propose a criterion (Figure 2), based on Palm's framework, to classify mathematical tasks as to be real or not. In this direction, to evaluate the presence of real tasks in the INVALSI tests, we started categorizing each task of every INVALSI test in terms of its feasibility. Feasibility, in accordance with Palm (2006) is made by two aspects: *event* and *question*. Event refers to the possibility that the event described in the school task can take, or not, place in real-life. Question refers to the possibility that the posed question might be posed, or not, in a corresponding real-life event. Those tasks that positively reflected both the aspects of event and question had been categorized as real. Then, for those real tasks, their quality was evaluated. Quality refers to the degree by which a task is similar to a corresponding real-life event. Specifically, quality is defined considering five aspects from Palm's framework, namely *existence*, *realism*, *specificity*, *presentation*, *language*. For each aspect a score was given: 0 if that aspect had no correspondence in a real-life event; 1 if that aspect had a partial correspondence in a real-life event; 2 if that aspect had complete correspondence in a real-life event. In conclusion, to every real task was associated its level of quality by calculating the mean among the scoring (0, 1 or 2) given to its indicators (*existence, realism, specificity, presentation, language*).

To answer to our second research question, we analyzed data from each test using the Rasch model (Rasch, 1960) which is the one adopted also by INVALSI statistical team. The Rasch model is a simple logistic model belonging to Item Response Theory and it provides a joint estimate of the difficulty of the items of a test and of the ability of each student by placing them on the same scale from −4 to +4. Specifically, in this study we considered a parameter measuring the difficulty for each

[1] More precisely the sample was composed by 43586 students in 2009, 35567 in 2010, 31564 in 2011, 30870 in 2012, 24774 in 2013 and 25349 in 2014.

item (Delta) and a parameter measuring the fit of the item (Weighted). Other information could be identified observing specific graphs called Distractor Plots, output of the Rasch Model. The Distractor plot of an item allows the comparison between the Item Characteristic Curve (ICC, which expresses the probability of responding correctly to a specific item depending on the difficulty of the item itself and the ability of the respondent) and the empirical trend of each possible answer, as function of students' ability. Moreover, the software used to implement the Rasch Model (ConQuest) gives a parameter, belonging to the Classical Test Theory, which highlights how much each item discriminates between students with high ability levels and students with lower ones. All the items classified as real were then classified on the bases of difficulty (Delta), fit of the item with the Rasch model (Weighted), and discrimination. This analysis from a psychometrical perspective is performed at an item level: a single mathematical task might be composed by 1 or more items, and each item is identified by a request. Furthermore, we also considered the Distractor plot of each of these items and started investigating possible features such as under-discrimination/over-discrimination, particular distractor trends (for example humped performance trend - Ferretti, Lemmo & Giberti, 2018) but also guessing effect[2] (correct answers given regardless of being certain about its trueness).

FEASIBILITY		
	Yes	No
Event	The event described in the school task can take place in real-life	The event described in the school task can not take place in real-life
Question	The question posed in the school task might be posed in a corresponding real-life event	The question posed in the school task might not be posed in a corresponding real-life event

(Yes; Yes)

QUALITY			
	0	1	2
Existence	Data and information available in the school task do not correspond to the ones available in a corresponding real-life event	Data and information available in the school task partially correspond to the ones available in a corresponding real-life event	Data and information available in the school task correspond to the ones available in a corresponding real-life event
Realism	Values given in the school task are not close to values in a corresponding real-life event	Values given in the school task are partially close to values in a corresponding real-life event	Values given in the school task are close to values in a corresponding real-life event
Specificity	Specifications of the school task are not analogous to the ones in a corresponding real-life event	Specifications of the school task are partially analogous to the ones in a corresponding real-life event	Specifications of the school task are analogous to the ones in a corresponding real-life event
Presentation	The task is not presented and formulated in a manner analogous to reality	The task is partially presented and formulated in a manner analogous to reality	The task is presented and formulated in a manner analogous to reality
Language	The language is not analogous to the language that might be used in a corresponding real-life event	The language is partially analogous to the language that might be used in a corresponding real-life event	The language is analogous to the language that might be used in a corresponding real-life event

Figure 2: Classification of mathematical tasks

Results

In Table 2 results from the classification of tasks as real and in terms of their quality are reported. The distributions of real tasks vary between 24% and 45%. The quality of the real tasks was relatively high in every test. These results give some first suggestions concerning the theoretical framework on which the test is constructed and the type of mathematics that the test is going to assess. Indeed, the presence of a considerable number of real tasks, all of them of medium-high quality, shows how INVALSI tests stressed the importance to give students opportunities to face with real-life problems, in accordance with the definition of mathematical competence and the Italian National Guidelines.

[2] Indeed if we consider multiple choice tasks with 4 possible answers the probability for a student to answer correctly randomly is 25% but we have information regarding the intensity of this phenomenon considering the intersection between the Rasch model and the y axes of the distractor plot.

Table 2: Results from the classification of real tasks

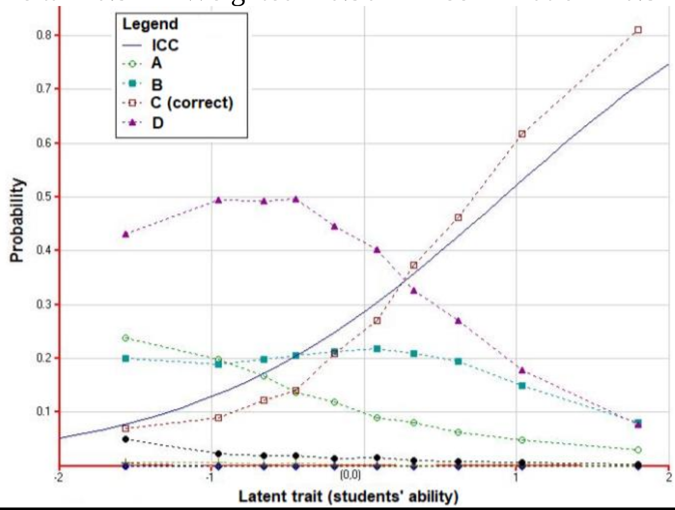
Year	Distribution	Total average of real tasks quality	SD
2008-09	24% (7/29)	1.4	0.3
2009-10	28% (9/32)	1.5	0.5
2010-11	33% (10/30)	1.7	0.3
2011-12	33% (11/33)	1.6	0.4
2012-13	31% (11/35)	1.7	0.3
2013-14	45% (13/29)	1.5	0.5

Then we classified on the bases of their psychometric features the 61 tasks (corresponding to 96 items in total) identified as real tasks. Considering items' difficulty (Delta parameter) we found that realistic tasks seem not to be a source of difficulty for students: the 35% of the items are easy with a delta smaller than -1; the 33% of the items shows a medium-low level of difficulty (-1<delta<0); the 27% of the items shows a medium-high level of difficulty (0<delta<1); and finally, we found only 4 difficult items (delta >1). The difficulty average is -0.61 highlighting that generally students are not particularly in trouble with real tasks. Furthermore, the item fit is considered good (values ranging from 0.8 and 1.2) for all the items except one, and the vast majority (84%) of the items identified have an optimal fit with values ranging between 0.9 and 1.1. Therefore, we can state that in the items identified the real context does not create a misfit with the model. The items identified as real are almost all discriminative items: the 81% have a good discrimination index (>0.3) and more than the 50% of the items have an optimal capacity to discriminate between students with lower ability levels and students with higher ones.

The items identified cover all item typology and all content domains but almost half of them belongs to the domain Data and Probability and only 5 items belong to the domain Space and Shapes, even if many real situations could become the stimulus for geometrical real problems.

In this research we started also to consider distractor plots of the identified real tasks, to detect possible recurring students' behavior facing a real task as function of their mathematical ability. This analysis will be the focus of further studies, but we report here an example of item analysis to explain its potentialities to investigate this issue. The task reported in table 3 was administered in the INVALSI grade 5 test in 2010. Accordingly to our proposed criterion, the task was classified as real with a quality of 1.2 (existence=1, realism=2, specificity=1, presentation=1, language=1). From a psychometrical perspective the item is medium-difficult with a difficulty index equal to 0.91 and the 32% of correct answers. The weighted index (0.9) highlights a good fit of the item. However, the distractor plot shows a little underdiscrimination of the item: the model overestimates the percentage of correct answers for students with medium-low ability and it underestimates it for higher ones. This characteristic might be linked to the influence of the real context that could give more difficulties for students of medium low ability levels due to the fact that they are not used, in the Italian school routine, to face numerous real problems. Moreover, also the trend of the incorrect answers is particularly interesting: distractor A decreases with students ability, this option might be chosen by students that find the correct answer “No” but they don’t link this answer to the correct argumentation; distractor B is constant for almost all ability levels; finally distractor D has an humped performance trend and is more attractive for students with medium-low ability levels. The guessing effect (probability for students of lower ability levels of choosing the correct answer randomly) is low as highlighted by the distractor plot.

Table 3: Example of a real INVALSI task (Task D14 grade 5 INVALSI test administered in 2010)

<table>
<tr><th>Task</th><th>Statistical data</th></tr>
<tr><td>Sandro has 20 dm of string to close four packages he needs to ship. For each package he needs 60 cm of string. Will he be able to close the four packages?
A. No, because 60 is greater than 20
B. Yes, because 20 dm are more than 6 dm
C. No, because 240 cm are more than 20 dm
D. Yes, because decimeters are larger than centimeters</td><td rowspan="3">Delta = 0.91 Weighted = 0.90 Discrimination = 0.51

</td></tr>
<tr><td>Our classification</td></tr>
<tr><td>Real task
Quality: 1.2</td></tr>
</table>

Conclusions

In this paper we reported the first results of an ongoing research project whose overall aim consists in studying what mathematical competencies and types of reasoning are intended to be assessed in the Italian standardized tests. Specifically, we focused on the presence of real tasks in INVALSI tests and on how students face with this kind of tasks. In order to find answers to our research questions, we firstly proposed a criterion, based on Palm’s framework (Palm, 2006) to classify real tasks. Results show that there is a consistent presence of real tasks in INVALSI tests. This fact suggests that one of the competences that these tests are intended to assess is the ability to develop and apply mathematical thinking to solve problems from everyday situations, as desired by the European Council and PISA. Concerning the psychometrical features, from a preliminary analysis of real task at an item level we can observe that, despite the limited use of real problems in didactical practices in Italy (Passarella, 2021), students do not reveal difficulties in these items, and most of them have a low or medium low difficulty level (Delta). This lack of work on real problems does not also influence the fit of these items except in a few cases. Indeed, real items are coherent with the latent trait measured by the whole test. An interesting outcome is that most of the real tasks are discriminative: students with high ability pass brilliantly these tasks, while students with low ability struggle more. Finally, we observe that in INVALSI test there is only a limited number of real tasks in geometry (Space and Shape), while this content domain could be one of the main interesting to promote this kind of tasks.

Future research is needed, particularly on: extending the analysis to tests up to 2019-20 and to the other grades of schooling, since in this study the focus was only on the fifth-grade; deepening the analysis of students' behaviors in facing real problems with some interviews to students and focus groups with teachers; investigating which other mathematical competencies are assessed in INVALSI tests; conducting a comparison with the PISA results in order to uncover and discuss differences or similarities between the two standardized tests.

References

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), 2895-2906. doi.org/10.29333/ejmste/90988

Giberti, C., & Maffia, A. (2020). Mathematics educators are speaking about PISA, aren't they?. *Teaching Mathematics and its Applications, 39*(4), 266-280. doi.org/10.1093/teamat/hrz018

MIUR (2012). *Indicazioni Nazionali per il Curricolo della Scuola dell'Infanzia e del Primo Ciclo di Istruzione*. www.indicazioninazionali.it/wp-content/uploads/2018/08/Indicazioni_Annali_Definitivo.pdf

NCTM (2000). *Principle and Standard for School Mathematics*. Va, NCTM. www.nctm.org/Standards-and-Positions/Principles-and-Standards/

Niss, M. (1992). Applications and modeling in school mathematics-directions for future development, in Wirzup, I. and Streit, R. (Eds), *Developments in school mathematics education around the world, 3*, National Council of Teachers of Mathematics.

OECD (2019). *PISA 2018 Mathematics Framework*. In *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. doi.org/10.1787/13c8a22c-en

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris: OECD Publishing. doi.org/10.1787/9789264173125-en

Palm, T. (2006). Word problems as simulations of real-world situations: a proposed framework. *For the Learning of Mathematics, 26*(1), 42-47.

Passarella, S. (2021). Mathematics Teachers' Inclusion of Modelling and Problem Posing in their Mathematics Lessons: an Exploratory Questionnaire. *European Journal of Science and Mathematics Education, 9*(2), 43-56. doi.org/10.30935/scimath/10773

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

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic Mathematics Education, In S. Lerman (Ed), *Encyclopedia of Mathematics Education* (pp. 521-525). Springer. DOI 10.1007/978-94-007-4978-8

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM*, *52*, 1-16. doi.org/10.1007/s11858-020-01130-4

Wiggins, G. (1993). Assessment: authenticity, context and validity, *Phi Delta Kappan, 75*(3), 200-214.