Percentages versus Rasch estimates: alternative methodological strategies for replication studies in mathematics education

Percentages versus Rasch estimates: alternative methodological strategies for replication studies in mathematics education

Clelia Cascella [a], Chiara Giberti[b]* and Andrea Maffia[c]

[a] *School of Environment, Education and Development, University of Manchester, Manchester, UK - clelia.cascella@manchester.ac.uk - ORCID 0000-0002-7735-5040;* [b]*Department of Human and Social Science, University of Bergamo, Bergamo, Italy - ORCID 0000-0001-7446-6709;* [c]*Department of Mathematics, University of Pavia, Pavia, Italy - ORCID 0000-0003-0080-1089*

*chiara.giberti@unibg.it

The authors are listed alphabetically, all jointly created this work and identified its aims and scopes. In addition, Chiara Giberti and Andrea Maffia collected primary data and developed the theoretical framework within which the results were interpreted. Clelia Cascella developed both the methodological and the analytical strategy and was responsible for data curation and analysis. All the authors contributed to the original writing of the manuscript, discussed the results from the analysis and have read and agreed to the current version of the manuscript.

Percentages versus Rasch estimates: alternative methodological strategies for replication studies in mathematics education

We present an external close replication of the 1985 Fischbein and colleagues' study about intuitive models of multiplication and division. We administered two batteries of mathematics items developed in the original study, via a spiralling process, to a quota sample of 903 students attending grade 7. Compared with the analytic strategy based on the count of correct answers employed in the original research, our study goes a step further as we propose a methodological approach that guarantees measurement invariance, thus allowing for the direct comparison of different groups of students and/or items. The advantages of Rasch estimates compared to percentages of correct answers over the total are critically discussed to show why the former should be considered as more robust than the latter.

Introduction

Many authors denounced the emergence of a crisis in science and, in particular, a crisis in replication studies at the beginning of the 21st century (Fanelli, 2018). Such crisis arises from the awareness of possible false positives in some studies and on the production of findings that cannot be replicated. In a recent study, Frias-Navarro and colleagues (2020) see this crisis as an opportunity to develop scientific research in terms of quality in production and in research practices, also thanks to the improvement of statistical, methodological, and technical instruments.

Replicability of studies in education might be more complex than in other fields: each intervention must consider numerous auxiliary assumptions, and this might even lead to the impossibility of falsification of the experiment (Frias-Navarro et al., 2020). The impossibility in reproducing identical conditions of an experiment in the educational field (Schmidt, 2009), is confirmed also by Cai and colleagues (2018) for research in mathematics education, which involves students, teachers, and their interactions in the mathematics classroom.

Within mathematics education, Sanchez-Aguilar (2020) argued that, despite their considerable importance, replication studies remain scarce, and the field of mathematics education is not well prepared to be open to replication studies.

In this paper, we present an alternative methodological approach adopted in work-in-progress research in which we replicate an important study published in 1985 by Fischbein and colleagues. With 184 citations on CrossRef, 430 citations on ResearchGate, and 849 citations on Google Scholar[1], Fischbein and colleagues' (1985) study is surely one of the most cited works on intuitive models of multiplication and division. Even recent publications rely on the results presented by Fischbein and colleagues (e.g. Qu et al., 2021; Erickson & Lockwood, 2021). Since this study has been so important for the field, remaking its findings and testing their robustness appears relevant. However, it is hardly replicable because relevant information (e.g. the sampling strategy) was not reported in the original study. We realized an as close as possible replication of this study: starting from the information published in Fischbein and colleagues' works, we used the same materials developed by Fischbein and colleagues and administered them to a quota sample that, as discussed in the methodological section, mirrors the characteristics of the Italian students' population. Compared with Fischbein's study, our study goes further as we employed a methodological strategy, alternative to that adopted in the original study.

Agreeing with Sanchez-Aguilar (2020) in terms of the worth of replication studies, we believe that our research could be important in the educational field, in which replication studies are still infrequent and, in particular, in the mathematics education field, as generalization of the previous research findings.

In the current paper we use Fischbein and colleagues' materials and show how an alternative methodological strategy (based on Rasch analysis) can overcome some of the

limits of that used in the original study, thus contributing to the current knowledge by proposing the methodology itself and by refreshing results from their study while providing substantive evidence about their robustness.

Replication studies

Assuming that an exact replication study is impossible in education (Sanchez-Aguilar, 2020), our replication study tries to remain as close as possible to the original one.

Schmidt (2009) differentiates between direct replication and conceptual replication: the former refers to an exact replication of an experimental procedure, the latter "attempts to test the same fundamental idea or hypothesis behind the original study, but the operationalizations of the phenomenon, the independent and dependent variables, the type and design of the study, and the participant population may all differ substantially" (Frias-Navarro et al, 2020, p. 626). Hence, considering the continuum between these two possible basic notions of replication (Schmidt, 2009; Sanchez-Aguilar, 2020), our study is to be considered as a *close replication study* (an intermediate model). Indeed, compared with the original study, we used the same materials (with few modifications explained in the "Materials" section), the same kind of participants (Italian grade 7 students; lower secondary school – on average, 12 years old), and the same variables, but we adopted a different approach in constructing the sample and used both the original and a more sophisticated methodology in analysing the data.

Clements (2015) proposed a second type of classification for replication studies: we could consider our study an *extension study*, which "means using new data-gathered on a sample representative of a different population, or gathered on the same sample at a substantially different time, or both" (Clements, 2015, p. 327).

Our new data are collected 35 years after the original study. Since the aim of our replication study is to "generalize" (in the sense of Sanchez-Aguilar, 2020) Fischbein and

colleagues' results, our sample was larger and constructed controlling context-variables. As part of the replication study, we employed the same analytical strategy developed by Fischbein and colleagues (see the methodological section). Then, we employed an alternative methodological strategy within the framework of the Rasch analysis that overcomes some shortcuts of the Classical Test Theory (CTT; Maclean et al., 2005). In our replication study, we provide all the necessary information to ensure that our research is transparent in terms of sampling strategy, data collection, methodologies, and materials, and we provide an open-access dataset (Mendeley database: Giberti, 2022), accessible for further investigations (Schoenfeld, 2018). Hence, we make our study replicable.

The original study carried out by Fischbein and colleagues

Research about intuitive models of multiplication and division dates back several years (e.g. Vest, 1971) and a large amount of research on this topic was published during the years, testifying how the interest in elementary models of multiplication/division remained high in time for researchers in education. This is still true in the recent years, when a new discussion about the harmonization of measure and multiplication is proposed by several authors (Simon et al., 2018; Izsàk & Beckmann, 2019; Polotskaia & Savard, 2020). Furthermore, the debate about elementary models of multiplication informs and nourishes the discussions about the teaching of multiplication properties (e.g. Maffia & Mariotti, 2018) and advanced multiplicative structures (e.g. Erickson & Lockwood, 2021).

In their seminal study published in 1985, Fischbein and colleagues proposed several word-problems like "1 kilo of oranges costs 1500 lire[2]. What is the cost of 3 kilos?", asking students to indicate the arithmetical operation allowing them to find a solution (the

solution itself was not requested). They divided their sample (628 students attending grade 5, 7, and 9) in two groups and administered two different forms of the same questionnaire composed of 21 items each. The items were designed to test the following hypothesis:

> Each fundamental operation of arithmetic generally remains linked to an implicit, unconscious, and primitive intuitive model. Identification of the operation needed to solve a problem with two items of numerical data takes place not directly but as mediated by the model. The model imposes its own constraints on the search process. (Fischbein et al., 1985, p. 4).

For instance, the intuitive model associated with multiplication is repeated addition. If this is true, a multiplication having a multiplier unconceivable as 'times' of iterations (e.g. a non-integer multiplier) would violate the intuitive model and should result as more difficult than a multiplication that is more easily explainable in terms of the intuitive model. Thus, a word-problem like "1 m of suit fabric costs 15000 lire. How much does 0.75 m cost?" has a lower percentage of correct answers compared with the item about the price of the oranges abovementioned.

The major finding provided in their work is the fact that the items in which the students provide the higher percentage of correct answers confirm repeated addition as an intuitive model for multiplication, and partitive and quotative divisions as intuitive models for division. If multiplication is interpreted as repeated addition, then (i) the multiplier must be a whole number, and (ii) the product must be bigger than the multiplicand. In partitive and quotative divisions, the divisor (a whole number in the partitive case) must be smaller than the dividend. Fischbein and colleagues believe that "many of the difficulties children encounter when dealing with arithmetical concepts and operations can be explained in a similar fashion as arising from the conflict between

formal algorithmic structures and related tacit, uncontrolled, primitive models" (Fischbein et al., 1985, p. 15).

As far as the authors of this paper know, any direct replication of Fischbein and colleagues' study has not been yet realized, nonetheless, conceptual repetitions of Fischbein and colleagues' study are not rare in literature. Indeed, these studies cannot be considered as direct replications because: word-problems are changed with out-of-context number sentences (Graeber & Tirosh, 1990); different word-problems, including different numbers, are introduced (De Corte et al., 1988; Bell et al., 1989; Harel et al., 1994); many of them address a population of different age than the original study (De Corte, et al., 1988; Kouba, 1989; Harel et al., 1994). These studies are usually aimed at testing if similar error patterns are found with similar problems, but the "similarity" of the problems may vary considerably. Some of these studies are qualitative (Kouba, 1989; Graeber & Tirosh, 1990) and confirm that, for some students, the difficulty in associating the correct operation to a word-problem may depend on the intuitive models described by Fischbein and colleagues (1985). Quantitative conceptual replication studies confirm that word-problems violating the intuitive model are more difficult, but they all rely on CTT (e.g. Bell et al., 1989; Harel et al., 1994), which is a limit as we will explain in the next sections.

Surely, conceptual replications provide important hints for generalization of results when confirmations are found; thus, Fischbein's description of the intuitive models of multiplication and division could be considered as a *solid finding* in the sense that these "results […] do not stand alone but have emerged from a line of research consisting of a larger set of related studies" (Dreyfus, 2017, p. 58). However, considering the strong differences of the mentioned studies, we wonder if drawing general conclusions about children behaviour, like "They seem to be a *natural outcome* of years

of work with whole numbers" (Graeber & Tirosh, 1990, p. 586, emphasis added), is appropriate on the base of the available data.

Direct replication studies are lacking in literature, and we believe that, given the importance of those results, it is particularly relevant to have replications to test their robustness. Also, according to modern research methods in the educational field, the study by Fischbein and colleagues (1985) is susceptible to updating. For instance, Fischbein and colleagues describe their sample as made of 628 pupils (228 in grade 5, 202 in grade 7 and 198 in grade 9) coming from 13 different schools in the same city, Pisa, in Italy. Neither information about the sample composition nor sampling criteria were provided. The lack of information about the sampling strategy makes the comparison of findings difficult, as it is well known that the sample characteristics can affect the outcome of educational research (e.g. Arnup et al, 2013; Baird, 2012). Then, considering possible important differences in the sample, our results could differ from those of the original study. Hence, we state our first research question as follows:

(1) Does the methodological approach employed here confirm the results of the Fischbein and colleagues' study?

Furthermore, Fischbein and colleagues' instruments are available in literature both in the original Italian version (Deri et al., 1983) and in English (Fischbein et al, 1985). The rationale behind the choice of the questions is explained, but the authors do not provide any measure of the validity and reliability of the instruments. The original authors explain that the total number of 42 items of the questionnaire was divided in two forms (21 items each) to reduce the fatigue effect. Each student responded to just one form. No evidence about the interchangeability of the two forms was provided in the original study, and so we are not sure if the comparison between items coming from different forms, administered to different students, is legitimate. To overcome this limit, in the current

study we equated the administered forms by concurrently calibrating them within the framework of the Rasch analysis (see the next paragraph), thus allowing for the direct comparability of the answers given by different students to different items (Kolenn & Brennan, 2014).

Furthermore, in our study, the Rasch model has also been used to validate the instrument administered to the students. Both equating and validation are not part of the replication study, but we conceive them as fundamental steps in our replication process as, in this way, we provided new insights onto the instrument used. Such an analytical strategy allows us to answer the following research question:

(2) Which are the psychometrical features of Fischbein et al.'s instrument? In particular: Is the instrument valid and reliable? How are the items distributed in terms of their difficulty in the two versions of Fischbein et al.'s forms?

Methods

Materials

In the original study, Fischbein and colleagues (1985) constructed a questionnaire composed of 42 word-problems, i.e. mathematical problems where important information is reported as a short narrative (Daroczy et al., 2015): the focus of the study were 12 multiplication and 14 division word-problems, but 16 addition/subtraction word-problems were also included in the questionnaire to reduce guessing effects. Starting from the full questionnaire, two forms (named A and B, respectively) were constructed and administered to two sub-groups of the sample. Each form was composed of 21 items, including 6 problems about multiplication, 7 about division, 8 about addition/subtraction. In the original study, authors provided the percentage of correct answers along with an indication of the most common error(s) for each item.

We started from the same two forms of the original study, in particular we considered all the questions already published in the Italian version of the questionnaire (Deri et al., 1983): this allowed us to be sure that no modification in question formulation was due to the translation of the text. We slightly modified some questions to replace words no longer of common use in Italian, nowadays, that thus could have hindered students' comprehension. For example, we substituted the Italian word '*ghisa'* with '*acciaio'* because, even though both the words refer to iron alloys, the former is less used in the common language (De Mauro & Chiari, 2016). Finally, we decided to avoid abbreviations for all the units of measure (e.g. we wrote 'kilograms' instead of 'kg').

The composition of the multiplication and division word-problems in the two forms was the same as in the original study. For the purpose of using a test equating technique to compare students' answers to the two forms, we created an anchor test (i.e. a subset of common items, identical in the two forms, and representing the mathematical content of the whole test) (Kolen & Brennan, 2014). We developed an anchor composed by the 8 addition/subtraction word-problems (from the form B of the original study) along with 4 items (2 multiplication problems and 2 division problems) that we developed by using similar formulation and the same numbers used in other problems.

A pilot study was then conducted to evaluate possible other ambiguities in the two forms of the test. The pilot involved 152 grade 7 students from 6 different schools: 74 students answered to form A and 78 students answered to form B. The results of the pilot, discussed with the teachers of the classes involved, highlighted that three other words were not clear for many students. Finally, we deleted one pronoun in one question and repeated the corresponding noun to be clearer.

Administration procedure

The two forms were embedded in SurveyMonkey and administered at the end of the

school year 2018/2019. The links were shared with the schools asking the teachers to assign randomly one version to half of the students in each classroom and the second version to the other half. SurveyMonkey automatically collects the answers and teachers did not have access to students' answers once they completed the questionnaire. Students entered SurveyMonkey accessing the website and inserting a code (provided by the teacher). The teachers were asked to help the students to get online, but not to provide any support in answering. Each student worked individually by using a laptop or a tablet provided by the school.

The online test started with short general instructions: students were asked not to perform the actual calculation, but only to indicate the operation used to solve the problem. Also, in the same fashion of Fischbein and colleagues' study (1985), an example was provided (Figure 1).

In the original study, Fischbein and colleagues reported that, to reduce any order effect, each of the two forms was ordered in two different ways. According to them, no order effect was found but no evidence supporting such claim was provided. Therefore, we decided to automatically randomize the order of the questions at each administration. This choice should avoid any order effect while mixing the questions related to different arithmetical operations. In the following, questions will be numbered for the sake of clarity in reporting the findings; that numbering does not correspond to the order of the items in the forms, but to the numbering provided in the original study.

After the administration, each teacher was asked to fill a report indicating any technical difficulty or any anomaly in the procedure. None of them reported any problem. Each student had a maximum time of 45 minutes to answer the questions but, according to the teachers' report, the whole time was never needed.

Methodological approach

Psychometric theory offers two approaches in analysing test data: Classical Test Theory and Item Response Theory (IRT) (e.g. Primi, 2017). CTT focuses on the total test score, by computing for example frequency of correct responses (to indicate item difficulty) or frequency of responses (to examine distracters) (e.g., Impara & Plake, 1998).

In the present study, students' answers were analysed within the framework of the Rasch analysis (Rasch, 1960/1980). The Rasch model, used to estimate students' *ability*[3] in mathematics (i.e. the latent trait), describes the probability of giving a correct answer to a dichotomously scored item as a function of the so-called students' 'relative ability', that is students' ability compared with item difficulty, as specified in equation [1].

$$\Pr(X_{ni} = 1 \mid \beta_n; \delta_i) = \frac{e^{(\beta_n - \delta_i)}}{1 + (\beta_n - \delta_i)}$$ $\Pr(X_{ni}=1)$= student's probability of encountering an item successfully β_n=student ability δi=item difficulty	[1]

As with the CTT (employed in the original study), both person and item parameters (i.e. student ability and item difficulty) are based on the count of correct answers (called *sufficient statistics*), given to all the items by each student or by all students to each item (**Errore. L'origine riferimento non è stata trovata.**2). Sufficient statistics are thus the total scores, and it is assumed that all the information about student ability or item difficulty resides in the total score, by column and by raw respectively (Andrich & Marais, 2019): "The best estimate of the ability parameter for a person can be derived from his raw score only" (Rasch, 1960/1980, p. 76). Even though a sufficient statistic does not provide an exact value for the latent trait, it summarizes *all that is known* on which to base an estimate of that measure. Raw scores are thus the input data for the Rasch analysis (rather than the output of the analysis, as in the CTT).

Even though CTT and IRT look so similar, there are several arguments favouring IRT (Andrich & Marais, 2019). Here, we focus just on (possible) differences between items difficulty (and thus persons' ability) computed within the framework of the IRT and of the CTT, and the related topics such as *measurement precision*. In particular, within the framework of the CTT, the measurement precision is assumed to be equal for all individuals irrespective of their attribute level, i.e. without accounting for the pattern of responses given by each student to all the administered items: if a student correctly answers two difficult items or two easy items over ten administered items, his/her estimated ability is 2 in 10, the items' difficulty is not accounted for and thus it does not contribute to the measurement of either students' ability or items' difficulty. In contrast, in the framework of the IRT, it is taken the whole pattern of item-scores into account. Therefore, it may reveal (even very subtle) changes in individuals' 'ability' that could go unnoticed "if one uses the sum scores from CTT" which ignore the pattern of the item scores" (Jabrayilov et al., 2016, p. 560).

Results achieved by employing CTT or IRT can be similar. For example, as recently pointed out by Jabrayilov, Emons, and Sijtsma (2016), "IRT is superior to CTT, provided that tests contain, say, at least 20 items, but *in general the differences between the two methods are small*." (p. 568).

Another critical difference, especially for the purposes of the current research is measurement invariance. CTT has been severely criticised because item difficulty and/or person ability based on the count of correct answers are sample dependent and, therefore, results based on percentages cannot be generalised (Hambleton, 2000). In contrast, IRT modelling overcomes such a limitation by estimating test-free persons' parameter and sample-free items' parameters (Wright & Stone, 1979): Rasch estimates are (i) person-free (Schimdt & Embretson, 2003) as item-parameters are independent from the students'

sample (in fact, as long as items are from the same pool, their estimated parameters do not change when items are administered to different samples); and, (ii) item-free (Schimdt & Embretson, 2003) as person-parameters are independent from items' characteristics thus allowing the comparability of their locations along the latent traits even when they were administered different items. Such a property, named *measurement invariance* (Engelhard, 2009; 2013), allows for comparing groups of students, groups of items, or groups of students with groups of items.

Rasch estimates are invariant measures of the latent trait only when data meet the model's assumptions: (i) unidimensionality, (ii) local independence; and (iii) cumulativity. In this paper, we tested data-model fit by using the traditional fit statistics that are standardized (Zstd) and mean-square (MNSQ) infit and outfit. Both infit and outfit are mean-square fit statistics. Their expected value is 1.0 with tolerable standard deviations around 0.20 (Engelhard, 2009).

Nonetheless, even if the Rasch model's assumptions hold, when both the test forms and the students sample change, the test forms may differ somewhat in difficulty, thus resulting in a biased comparison between groups of students and/or groups of items (Kolen & Brennan, 2014). Within the framework of the IRT modelling, a number of techniques have been developed to equate students' answers so that both person- and item-parameters could be expressed in the same (*logit*) metric. Converting test metric (of each single form) onto a common metric is necessary to (i) use different questionnaire-forms interchangeably; and (ii) put all items (administered in different forms) and all students (belonging to different groups) onto the same latent trait, thus allowing for robust comparisons between sub-groups of students and/or items. Skipping test forms equalization may result in biased estimates and hinder a meaningful comparison of

different forms. To this end, we performed a concurrent calibration (Kolen & Brennan, 2014) in Winstep (Linacre, 2022).

Then, according to previous literature (e.g., Andrich & Marais, 2019; Kolen et al., 2014), we claim that the Rasch model has to be preferred at least for three reasons: first, Rasch estimates are invariant across sub-groups of items and/or students, thus allowing for the generalization of results (measurement invariance); second, Rasch analysis provides a framework for the validation of the administered questionnaire-forms thus allowing us to go a step further compared with the original study; and, third, within the framework of the Rasch analysis, we can *equate* different questionnaire-forms and put results (both item difficulty and student ability) onto the same metric, thus making them directly comparable.

Sampling strategy

Since sampling criteria were not presented in the original study, we used those employed by the Italian national institute for the evaluation of the educational system (INVALSI). INVALSI is responsible for national assessment and thus administers every year its achievement tests to a sample statistically representative of the entire population. For the purposes of the present research, primary data were collected in Emilia-Romagna, a big region in northern Italy, where students' mathematical performances in the national assessments are in line with the national average (Figure 3). Ethics approval was granted by the Department of Mathematics of the University of Pavia. Participation in the study was voluntary and active parental consent was obtained for each student.

For forms equating, the ideal sample size within the framework of the Rasch analysis is 400 students per form (Kolen & Brennan, 2014). Our sample consists of 903 students, attending grade 7 (on average, 12-years old).

INVALSI collects data through a two-stage sampling design, stratified by school and region (Falorsi et al., 2019). In the attempt to replicate INVALSI sample structure, we used a proportional quota sampling (Moser, 1952), a non-random sampling design that requires defining groups (i.e., quotas) of units by using specific (non-random) criteria. The quota sampling strategy uses key categories in the larger population to specify how many members of the sample should fall into each of those categories or combinations of categories. This sampling is a non-probability technique because it requires only that the quota for each category is met without any further attention to how those sample members are located. Yang and Banamah (2014) claimed that quota sampling should be taken as an acceptable alternative to probability sampling. Indeed, proportional quota sampling has been extensively used as a quicker and cheaper alternative to random sampling because, when quotas are merged into the final sample, "the characteristics of the sample precisely match specified demographic characteristics of the survey population" (Lewis-Beck et al., 2004, p. 906).

Nonetheless, as with Given (2012), we acknowledge that quota samples are not truly generalizable because even when the quota-based categories match the corresponding categories in the target population, the sample may not represent other critical characteristics outside the quota system. Nonetheless, our research was not aimed to draw a statistically representative sample. Yet, we aimed to represent in our sample the characteristics of the real students' population under the assumption that the exploration of students' answers cannot disregard how they may change depending on students' personal characteristics (e.g., Cascella, 2020a), such as their socioeconomic status. Previous studies have shown, for example, that (i) low-SES (socioeconomic status) students develop attainment more slowly compared to students from higher SES groups (Sidanius & Pratto, 1999) and, in particular, that "on average, math scores of students

with indicators of high-socioeconomic status (SES) are over one standard deviation above those with low SES indicators" (Baird, 2012, p. 484). Moreover, previous studies have shown that aggregated SES (at classroom and/or school level) is as important as individual SES (e.g., Coleman, 1966). The association between aggregated SES and individual students' attainment increases over time, moving from primary to secondary school, and becomes more important than that between individual SES and attainment (Cascella, 2020b).

In the present study, we used school SES composition as the first key variable to identify the quotas: starting from the entire list of lower secondary schools located in Emilia-Romagna (provided by INVALSI), we stratified schools depending on the SES of their students.

To measure students' SES, we used a sociocultural index (SC-index) based on the combination of highest parental education and occupation (Cascella, 2019)[4]. Results based on data collected by INVALSI (from 2014/15 to 2018/19) showed that SC-index ran in [–3; +2], with mean 0.02 and standard deviation 0.86[5]. Based on the combination of SC-index's mean and standard deviation, we identified three groups of schools (our quotas), i.e. schools with (i) low, (ii) medium, and (iii) high school SES. Then, we calculated the percentage of schools belonging to each quota. The proportional quota sample constructed for the purposes of the present study looked like that constructed by INVALSI in the same academic year (Table 1).

Finally, in addition to students' SES, we looked at the distribution of students by students' (i) sex (males or females), (ii) citizenship status (native, first- and second-generation students), (iii) the number of classrooms per school, taken as an indicator of school size, (iv) the number of students per classroom, and (v) the proportion of regular, in advance, and retained students (Table 2). We considered these variables because

previous studies based on Italian data showed that they are significantly associated with attainment (e.g., INVALSI, 2018).

Results

Replication

In the work by Fischbein and colleagues (1985) the results of the multiplication and division items were provided in the form of percentage of correct answers.

In table 3, we compared our results with those from the original study in terms of percentage of correct answers. There are some tasks for which we can observe differences in the percentages, but these differences do not question the nature of Fischbein et al.'s results. For instance, the percentage of correct answers for question Q23B (I paid 900 lire for 0.75 hectogram of cocoa; what is the price of 1 hectogram?) is higher in our replication (41%) than the original study (25%). This partitive division problem is compared, in the original study, with Q13A and Q14B which are still partitive problems having a whole number as operator. Both in the original study as in our replication, Q13 and Q14 have much higher percentage of correct answer than Q23B; this supports the fact that, according to intuitive model of quotative division, the operator must be an integer (see Giberti & Maffia, 2022). In the original paper, Q23B is also contrasted with Q21A (still partitive division) since a decimal number is the operator in the former and the operand in the latter. In the original study, there is a strong difference in the percentage of correct answers, while this is not the case in our replication. This may depend on the fact that one of the effects of the intuitive model was mitigated by other factors, but it is also possible that the comparison between these two tasks was not appropriate in the case of Fischbein et al.'s study, because these tasks belong to two different forms which are not equivalent as noted below.

Rasch modelling

Data-model fit and questionnaire validation.

Before estimating person and item parameters, we double-checked data-model fit. All items showed both infit and outfit close to 1 (the ideal value), thus supporting the unidimensionality hypothesis (Table 4).

Further, we investigated both person- and item-separation and reliability that provide evidence about construct validity and for the sensitivity of our measure in separating people and items along the latent trait (Table 4).

Both item-separation and reliability were above the cut-off points (i.e., 3 and 0.9, respectively), thus providing empirical evidence of the reproducibility of measure location. Moreover, results indicated that the sample was large enough to confirm the item difficulty hierarchy of the instrument, thus supporting the construct validity of the instrument used in the present study.

Person-separation and reliability was slightly lower than the cut-off point (i.e., lower than 2 and 0.8, respectively) thus suggesting that the instrument may be not sensitive enough to properly scale subjects along the latent trait, i.e. to distinguish between high and low performers.

Both the CTT and Rasch model allows for ordering items (and subjects) depending on their difficulty (and relative ability). The Item-Person map (Figure 4) shows the distribution of both persons and items along the same latent trait. As expected, the items' order based on Rasch estimates and on the count of current answers are pretty much the same with a very few exceptions when we analyse our data, but slightly different when we compare results from our study with those from the original one[6].

Discussion and conclusion

We presented an external close replication (Sanchez-Aguilar, 2020) of a renowned study in mathematics education. Fischbein and colleagues' (1985) study about intuitive models of multiplication and division has inspired many researchers but can be updated from the methodological point of view.

Differently than the original study, we employed a proportional quota sampling and provided full details about the criteria employed to draw it up, thus making our study easily replicable. For answering our first research question, we started from the assumption (based on available literature) that the Rasch model should be preferred at least because Rasch estimates are invariant across sub-groups of items and/or students; Rasch analysis provides a framework for the validation of the administered questionnaire-forms; we can equate different questionnaire-forms and put results onto the same metric.

When we compared the count of correct answers in our data to the count realized by Fischbein et al. (1985), we observed some differences in the percentage of correct answers. However, we can state that our study confirms the results of the original one, because the effects of the intuitive model described in the original study are still predictive of the relative difficulty of the tasks[6].

It is worth noting that results of the current study based on the count of correct answers and those from the Rasch analysis are similar as the Rasch test scores are based on the sufficient statistics (i.e., the number of correct answers provided by all students to each item compared with the number of wrong answers). Results based on the Rasch analysis are more accurate than those based on the percentage, but they are consistent.

Thus, we can answer our first research question by stating that a more robust analytical approach confirmed the solidity (in the sense of Dreyfus, 2017) of the results from the original result.

Considering the second research question, our results showed that the forms developed in the original study are not equivalent, thus questioning the comparability of results in the original study and recommending, in future replication research, the administration of both forms – since they are not equivalent, it might not be possible to replicate the same result administering only one of the two forms.

In the present study, we anchored the forms and equated them. Making answers comparable is a critical methodological improvement compared with the original study as Fischbein and colleagues assumed the comparability of both the forms and the two groups of students the forms were administered to. Nonetheless, the authors assumed such comparability without providing any prove of the direct comparability of neither the tests nor the answers. Results from our investigations showed that A and B cannot be used interchangeably, as also suggested by the different items' hierarchy from the current and the original study (Table 5).

Moreover, within the framework of the Rasch analysis, we explored instruments' validity by investigating data-model fit and calculated both person- and item-separation and reliability. Then, we can answer to our second research question by stating that results suggest the necessity of deeper studies on the strength of the scale from a psychometric point of view. This is beyond the aims of a replication study but provides a possible line for future research. Also, further analysis and implementation of the questionnaires may help in selecting the items that are more representative of a certain phenomenon (namely of a certain intuitive model of multiplication or division) while preserving the reliability and validity of the instrument. That instrument would help in realizing more easily further replications, then deepening our understanding on how intuitive models intervene in the learning of arithmetic for different students. Such instrument may be used in studies about

the effectiveness of the introduction of different models of multiplication/division when these operations are introduced for the first time in primary school.

As a final remark, we add our voices to those researchers who have stressed the importance of continuing the replication of studies to strengthen the obtained results (Sanchez-Aguilar, 2020) and to provide solid findings to the field of mathematics education (Bolondi & Ferretti; 2021; Dreyfus, 2017; Primi, 2017). Then, we encourage the use of the data we are providing for further comparisons.

Endnotes

1. Sources: search.crossref.org - www.researchgate.net - scholar.google.com - data updated at 9th June 2022
2. 'Lire' was the name of the currency in Italy before the introduction of the Euro in 2002.
3. 'Ability' is a technical term used within the framework of the Rasch analysis to refer to the latent trait that the instrument used to collect data purports to measure. In this study, students' ability thus refers to their mathematical attainment.
4. A detailed description of the procedure employed to construct the SC-index has been provided in Cascella (2019) along with a description of the procedure to deal with missing values and data imputation.
5. INVALSI collects data at grade 2 and 5 (primary school), grade 8 (lower secondary), and at grade 10 and 13 (secondary school). No data are collected in any other grade. Nonetheless, since lower secondary school spans from grade 6 to grade 8, and since our sample units are schools (instead of students), information collected at grade 8 by INVALSI can be considered reliable to acquire the information we need (to construct the SC-index).
6. The presence of differences between the results of the original study and our replication could depend on many causes (included the different samples, the changes in Italian curriculum during the last 35 years, and so on) that could be researched. In any case, the available data do not allow us to conjecture any interpretation and questioning the nature of those differences is beyond the scopes of this paper.

Disclosure statement

The authors report there are no competing interests to declare.

References

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. *Measuring in the Educational, Social and Health Sciences*, 41–53.

Arnup, J. L., Murrihy, C., Roodenburg, J., & McLean, L.A. (2013). Cognitive style and gender differences in children's mathematics achievement. *Educational Studies, 39*(3). 355–368. https://doi.org/10.1080/03055698.2013.767184

Baird, K. (2012). Class in the classroom: The relationship between school resources and math performance among low socioeconomic status students in 19 rich countries. *Education Economics, 20*(5), 484–509. https://doi.org/10.1080/09645292.2010.511848

Bell, A., Greer, B., Grimison, L., & Mangan, C. (1989). Children's performance on multiplicative word problems: Elements of a descriptive theory. *Journal for Research in Mathematics Education, 20*(5), 434–449. https://doi.org/10.5951/jresematheduc.20.5.0434

Bolondi, G., & Ferretti, F. (2021). Quantifying Solid Findings in Mathematics Education: Loss of Meaning for Algebraic Symbols. *International Journal of Innovation in Science and Mathematics Education, 29*(1), 1–15. https://doi.org/10.30722/IJISME.29.01.001

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., & Hiebert, J. (2018). The role of replication studies in educational research. *Journal for Research in Mathematics Education, 49*, 2–8. https://doi.org/10.5951/jresematheduc.49.1.0002

Cascella, C. (2019). How much 'home possession' affect educational attainment? Empirical evidence towards a simpler socio-economic status index. In L. Gómez Chova, A. López Martínez, I. Candel Torres (Eds.) *ICERI2019 Proceedings* (pp. 5809-5814). IATED.

Cascella, C. (2020a). Exploring the complex relationship between students' reading skills and their performance in mathematics: a population-based study. *Educational Research and Evaluation, 26*(3-4), 126–149. https://doi.org/10.1080/13803611.2021.1924790

Cascella, C. (2020b). Intersectional effects of Socioeconomic status, phase and gender on Mathematics achievement. *Educational Studies, 46*(4), 476–496. https://doi.org/10.1080/03055698.2019.1614432

Clements, M. A. (2015). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys, 31*(1), 326–342. https://doi.org/10.1111/joes.12139

Coleman, J. S. (1966). *Equality of educational opportunity* [summary report] (Vol. 1). US Department of Health, Education, and Welfare, Office of Education.

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. https://doi.org/10.3389/fpsyg.2015.00348

De Corte, E., Verschaffel, L., & Van Collie, V. (1988). Influence of number size, problem structure, and response mode on children's solution of multiplication problems. *Journal of Mathematics Behaviour, 7*, 197–216.

De Mauro, T., & Chiari, I. (2016). Il Nuovo vocabolario di base della lingua italiana. *Internazionale*. Retrieved from: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

Deri, M., Nello, M., & Marino, M. (1983). Il ruolo dei modelli primitivi per la moltiplicazione e la divisione. *L'insegnamento della matematica e delle scienze integrate, 6*(6), 6–27.

Dreyfus, T. (2017). What are solid findings in mathematics education? In Dooley, T. & Gueudet, G. (Eds.) *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (pp. 57-62). DCU Institute of Education & ERME.

Engelhard, G. (2009). Item and person functioning for Students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602. https://doi.org/10.1177/0013164408323240

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge. https://doi.org/10.4324/9780203073636

Erickson, S.A., & Lockwood, E. (2021). Investigating Combinatorial Provers' Reasoning about Multiplication. *International Journal of Research in Undergraduate Mathematics Education, 7*, 77–106. https://doi.org/10.5951/jresematheduc-2021-0112

Falorsi, P. D., Falzetti, P. & Ricci, R. (2019). *Le metodologie di campionamento e di scomposizione della devianza nelle rilevazioni nazionali dell'INVALSI*. Il Mulino.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *PNAS, 115*, 2628–2631. https://doi.org/10.1073/pnas.1708272114

Fischbein, E., Deri, M., Nello, M., & Marino, M. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 16*(1), 3–17. https://doi.org/10.5951/jresematheduc.16.1.0003

Frias-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perezgonzalez, J., & Berrios-Riquelme, J. (2020). Replication crisis or an opportunity to improve scientific production? *European Journal of Education, 55*(4), 618–631. https://doi.org/10.1111/ejed.12417

Giberti, C. (2022). *NewData_FischbeinIntuitiveModels.* Mendeley Data - V3. doi: 10.17632/4gmd9xnwhs.3

Giberti, C., & Maffia, A. (2022). Primitive model of partitive division: A replication of the Fischbein and colleagues' study. *Implementation and Replication Studies in Mathematics Education, 2*(2), 149–173. https://doi.org/10.1163/26670127-bja10007

Given, L. (2012). Quota sampling. *The SAGE Encyclopedia of Qualitative Research Methods, 2008*, 1–3.

Graeber, A. O., & Tirosh, D. (1990). Insights fourth and fifth graders bring to multiplication and division with decimals. *Educational Studies in Mathematics, 21*(6), 565–588. https://doi.org/10.1007/BF00315945

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical care*, II, 60-65.

Harel, G., Behr, M., Post, T., & Lesh, R. (1994). The impact of the number type on the solution of multiplication and division problems: further investigations. In G. Harel & J. Confrey (Eds.) *The Development of Multiplicative Reasoning in the Learning of Mathematics* (pp. 365–384). State University of New York Press.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69-81.

INVALSI (2018). *Rapporto risultati* [Tr. Results Report]. INVALSI.

Izsák, A., & Beckmann, S. (2019). Developing a coherent approach to multiplication and measurement. *Educational Studies in Mathematics, 101*(1), 83–103. https://doi.org/10.1007/s10649-019-09885-8

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/01466216166640

Kolen, M. J., & Brennan, R.L. (2014). *Test equating, scaling, and linking*. Springer.

Kouba, V. L. (1989). Children's solution strategies for equivalent set multiplication and division word problems. *Journal for Research in Mathematics Education, 20*(2), 147–158. https://doi.org/10.5951/jresematheduc.20.2.0147

Lewis-Beck, M. S., Bryman, A., & Futing Liao, T. (2004). *The SAGE encyclopedia of social science research methods* (Voll. 1–0). Sage Publications, Inc.

Linacre, J.M. (2022). Winsteps® (Version 5.2.3) [Computer Software]. Available from https://www.winsteps.com/

Maclean, R., Watanabe, R., Baker, R., Boediono., Cheng, Y. C., Duncan, W., Keeves, J., Mansheng, Z., Power, C., Rajput, J. S., Thaman, K. H., Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *Applied Rasch Measurement: A Book of Exemplars*. Springer. https://doi.org/10.1007/1-4020-3076-2

Maffia, A., & Mariotti, M.A. (2018). Intuitive and formal models of whole number multiplication: Relations and emerging structures. *For the Learning of Mathematics, 38*(3), 30–36. https://www.jstor.org/stable/26548509

Moser, C. A. (1952). Quota sampling. *Journal of the Royal Statistical Society, 115*(3), 411–423.

Polotskaia, E., & Savard, A. (2020). Some multiplicative structures in elementary education: a view from relational paradigm. *Educational Studies in Mathematics, 106*(3), 447–469. https://doi.org/10.1007/s10649-020-09979-8

Primi, C. (2017). Solid findings in mathematics education: A psychometric approach. In Dooley, T. & Gueudet, G. (Eds.) *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (pp. 63-67). DCU Institute of Education & ERME.

Qu, C., Szkudlarek, E., & Brannon, E.M. (2021). Approximate multiplication in young children prior to multiplication instruction. *Journal of Experimental Child Psychology, 207*, 105–116. https://doi.org/10.1016/j.jecp.2021.105116

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research* (Expanded edition, 1980). University of Chicago Press.

Sanchez-Aguilar, M. (2020). Replication Studies in Mathematics Education: What Kind of Questions Would Be Productive to Explore? *International Journal of Science and Mathematics Education, 18*, 37–50. https://doi.org/10.1007/s10763-020-10069-7

Schmidt, K.M., & Embretson, S.E. (2003). Item Response Theory and measuring abilities. In J. A. Schinka & W. F. Velicer (Eds.) *Handbook of psychology, 2, Research Method in Psychology* (pp. 429–445), John Wiley and sons.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/14805-036

Schoenfeld, A. H. (2018). On replications. *Journal for Research in Mathematics Education, 49*(1), 91–97. https://doi.org/10.5951/jresematheduc.49.1.0091

Sidanius, J., & Pratto, F. (1999). *Social Dominance*. https://doi.org/10.1017/cbo9781139175043

Simon, M. A., Kara, M., Norton, A., & Placa, N. (2018). Fostering construction of a meaning for multiplication that subsumes whole-number and fraction multiplication: A study of the Learning Through Activity research program. *The Journal of Mathematical Behavior, 52*, 151–173. https://doi.org/10.1016/j.jmathb.2018.03.002

Vest, F. (1971). A catalogue of models for multiplication and division of whole numbers. *Educational Studies in Mathematics, 3*(2), 220–228. https://doi.org/10.1007/BF00305450

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Yang, K., & Banamah, A. (2014). Quota sampling as an alternative to probability sampling? An experimental study. *Sociological Research Online, 19*(1), 1–11. https://doi.org/10.5153/sro.319

Table 1. Number of schools in each SES quota.

SES composition	Population		INVALSI sample		Our sample	
	N	%	N	%	N	%
Low	14957	40.1	660	40.3	6	40
Medium	10571	28.4	472	28.8	5	33
High	8065	21.6	391	23.9	8	27
Missing	*3657*	*9.8*	*115*	*7.0*	*0*	*0*
Total	37250	100.0	1638	100.0	15	100

Source: our elaboration on INVALSI data collected in 2017 at grade 8

Table 2. Students' characteristics by school.

	Gender		Foreign students			Regularity			
School	Boys	Girls	2nd generation	1st generation	Total	In advance students	Retained students	Classrooms in the school	Students per classroom (average)
1	53%	48%	7%	5%	12%	5%	1%		
2	49%	51%	0%	0%	0%	0%	0%	18	24
3	54%	46%	10%	8%	18%	0%	0%	6	27
4	53%	47%	0%	0%	0%	1%	0%	18	24
5	51%	49%	5%	3%	8%	1%	3%	18	24
6	58%	42%	-	-	7%	3%	2%	16	22
7	52%	48%	13%	5%	18%	3%	0%	18	23
8	55%	45%	-	-	17%	1%	1%	18	23
9	53%	47%	10%	8%	18%	1%	7%	12	21
10	49%	50%	18%	5%	23%	0%	3%	15	22
11	54%	46%	7%	4%	11%	-	6%	15	26
12	51%	49%	14%	2%	16%	4%	3%	21	25
13	53%	47%	15%	3%	18%	0%	0%	18	21
14	53%	47%	17%	3%	20%	0%	0%	65	21

Table 3. Items' difficulties from the Rasch analysis and those based on the percentage of correct answers. Each item is labelled with the letter Q followed by the numbering provided in the original study by Fischbein and colleagues. The letter A or B is added at the end of the code to indicate if the item belonged to form A or form B. For instance, item Q1A corresponds to the first item presented in Fischbein et al. (1985), while item Q26A is the last one, and they both belong to form A.

Order of items based on Rasch (current study)		Order of items based on percentage (current study)		Order of items based on percentage (original study)	
Item	Measure	Item	correct answers (%)	Item	correct answers (%)
Q13A	-2.75	Q13A	0.93	Q1A	0.96
Q2B	-2.08	Q2B	0.93	Q10B	0.93
Q14B	-1.54	Q14B	0.88	Q13A	0.93
Q10B	-1.41	Q10B	0.86	Q8A	0.91
Q15B	-1.25	Q15B	0.85	Q2B	0.91
Q1A	-0.86	Q1A	0.78	Q14B	0.90
Q8A	-0.80	Q8A	0.77	Q15B	0.89
Q19B	-0.61	Q19B	0.76	Q11A	0.85
Q26A	-0.51	Q26A	0.73	Q18A	0.79
Q12B	-0.28	Q12B	0.72	Q12B	0.78
Q3A	0.06	Q3A	0.65	Q21A	0.77
Q11A	0.13	Q11A	0.64	Q19B	0.77
Q18A	0.18	Q18A	0.63	Q22B	0.74
Q25B	0.47	Q25B	0.61	Q3A	0.74
Q9B	1.07	Q22B	0.48	Q20A	0.71
Q22B	1.21	Q9B	0.47	Q25B	0.63
Q21A	1.25	Q21A	0.46	Q26A	0.62
Q23B	1.37	Q20A	0.42	Q6A	0.57
Q24A	1.43	Q24A	0.41	Q4A	0.57
Q20A	1.49	Q23B	0.41	Q7B	0.43
Q4A	1.94	Q4A	0.33	Q24A	0.38
Q6A	2.23	Q6A	0.29	Q9B	0.38
Q5B	2.29	Q5B	0.28	Q17B	0.30
Q7B	3.05	Q17B	0.19	Q23B	0.25
Q17B	3.12	Q7B	0.18	Q6A	0.24
Q6A	3.32	Q6A	0.16	Q5B	0.18

Note. Items were ordered by difficulty, from the easiest to the most difficult. Differences in the order according to Rasch model and according to percentage of correct answers (results of the current study) are highlighted in grey.
Source: our elaboration

Table 4 Items fit statistics. Items are ordered by difficulty, from the easiest to the most difficult.

Item	Measure	Model S.E.	INFIT_MNSQ	OUTFIT_MNSQ
Q13A	-2.75	0.24	0.96	1.90
Anchor_item_1	-2.66	0.18	1.08	1.55
Anchor_item_2	-2.44	0.16	1.01	1.46
Anchor_item_3	-2.33	0.16	0.97	1.36
Anchor_item_4	-2.16	0.15	0.95	0.82
Q2B	-2.08	0.21	0.91	0.66
Anchor_item_5	-1.95	0.14	0.94	0.99
Anchor_item_6	-1.74	0.13	0.94	1.01
Q14B	-1.54	0.18	0.92	1.17
Q10B	-1.41	0.17	0.95	0.79
Q15B	-1.25	0.17	0.89	0.91
Anchor_item_7	-1.10	0.11	0.96	0.97
Anchor_item_8	-1.09	0.11	0.88	0.76
Q1A	-0.86	0.14	0.95	0.72
Q8A	-0.80	0.14	0.86	0.70
Q19B	-0.61	0.14	1.11	1.24
Q26A	-0.51	0.13	0.99	1.01
Q12B	-0.28	0.13	1.02	0.86
Q3A	0.06	0.12	0.96	1.27
Q11A	0.13	0.12	0.97	0.98
Anchor_item_9	0.14	0.09	0.94	0.89
Q18A	0.18	0.12	1.00	1.09
Anchor_item_10	0.31	0.08	1.03	1.00
Q25B	0.47	0.12	0.90	0.83
Anchor_item_11	0.90	0.08	1.03	1.16
Q9B	1.07	0.12	1.03	1.17
Q22B	1.21	0.11	1.12	1.15
Q21A	1.25	0.11	1.12	1.24
Q23B	1.37	0.12	1.05	1.07
Q24A	1.43	0.11	1.16	1.24
Q20A	1.49	0.11	1.00	1.04
Anchor_item_12	1.61	0.08	1.01	1.21
Q4A	1.94	0.12	0.97	1.06
Q6A	2.23	0.12	0.87	0.88
Q5B	2.29	0.13	0.93	1.31
Q7B	3.05	0.15	0.87	1.10
Q17B	3.12	0.15	1.19	1.17
Q16A	3.32	0.14	1.09	1.90

Note. PERSON: REAL SEP.: 1.82 REL.: 0.77; ITEM: REAL SEP.: 12.20 REL.: 0.99
Source: our elaboration

Giacomo has got 3 red pens and 4 blue pens, how many pens has he got all together? Correct format of the answer: 3+4

Figure 1. Example provided.

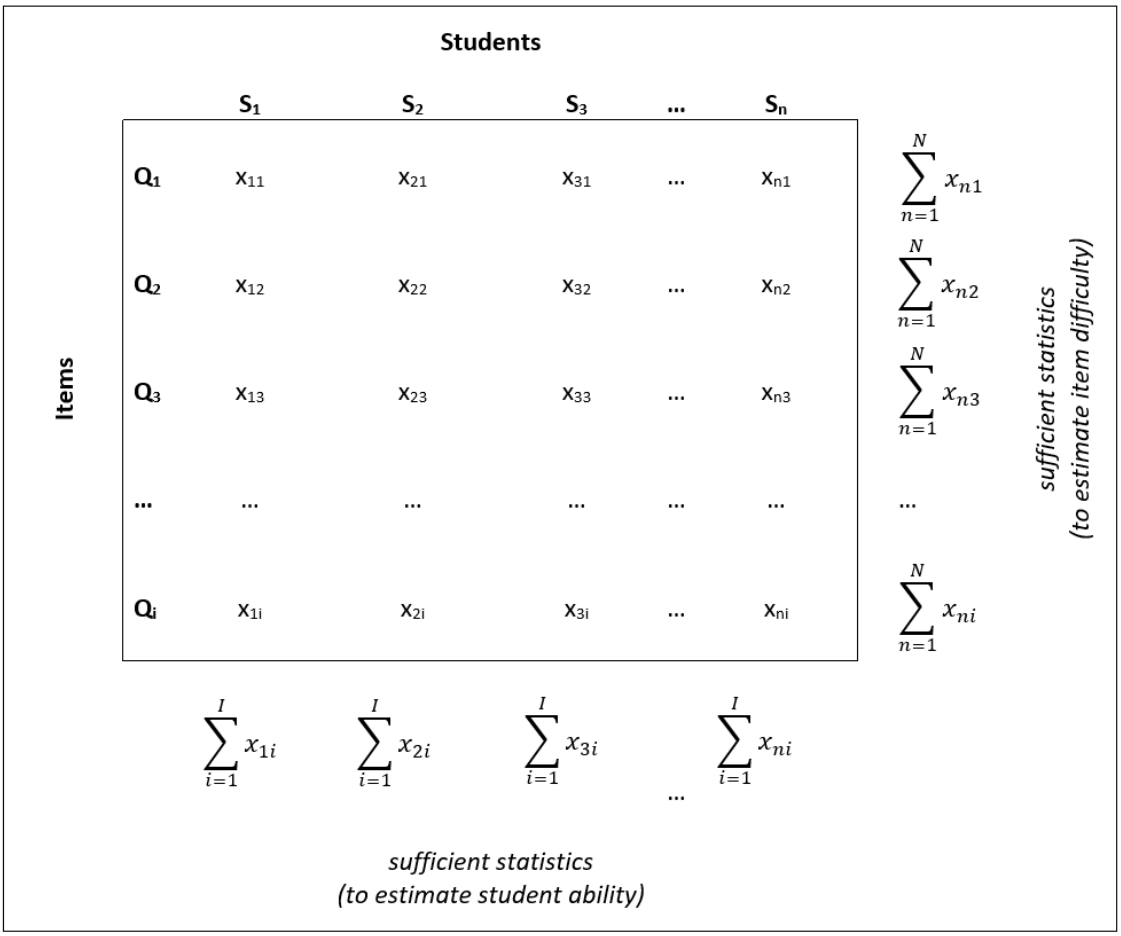

Figure 2. Rasch Sufficient Statistics (total scores to estimate student ability and item difficulty).

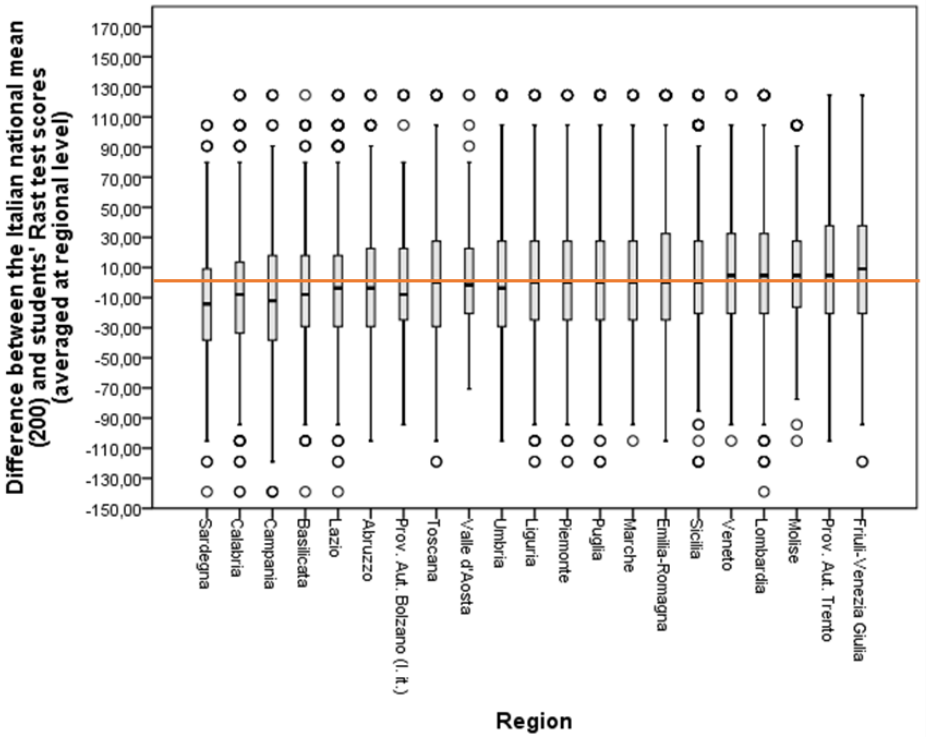

Figure 3. Deviation of students' Rasch test scores in mathematics from the Italian national mean in the different Italian geographical regions.

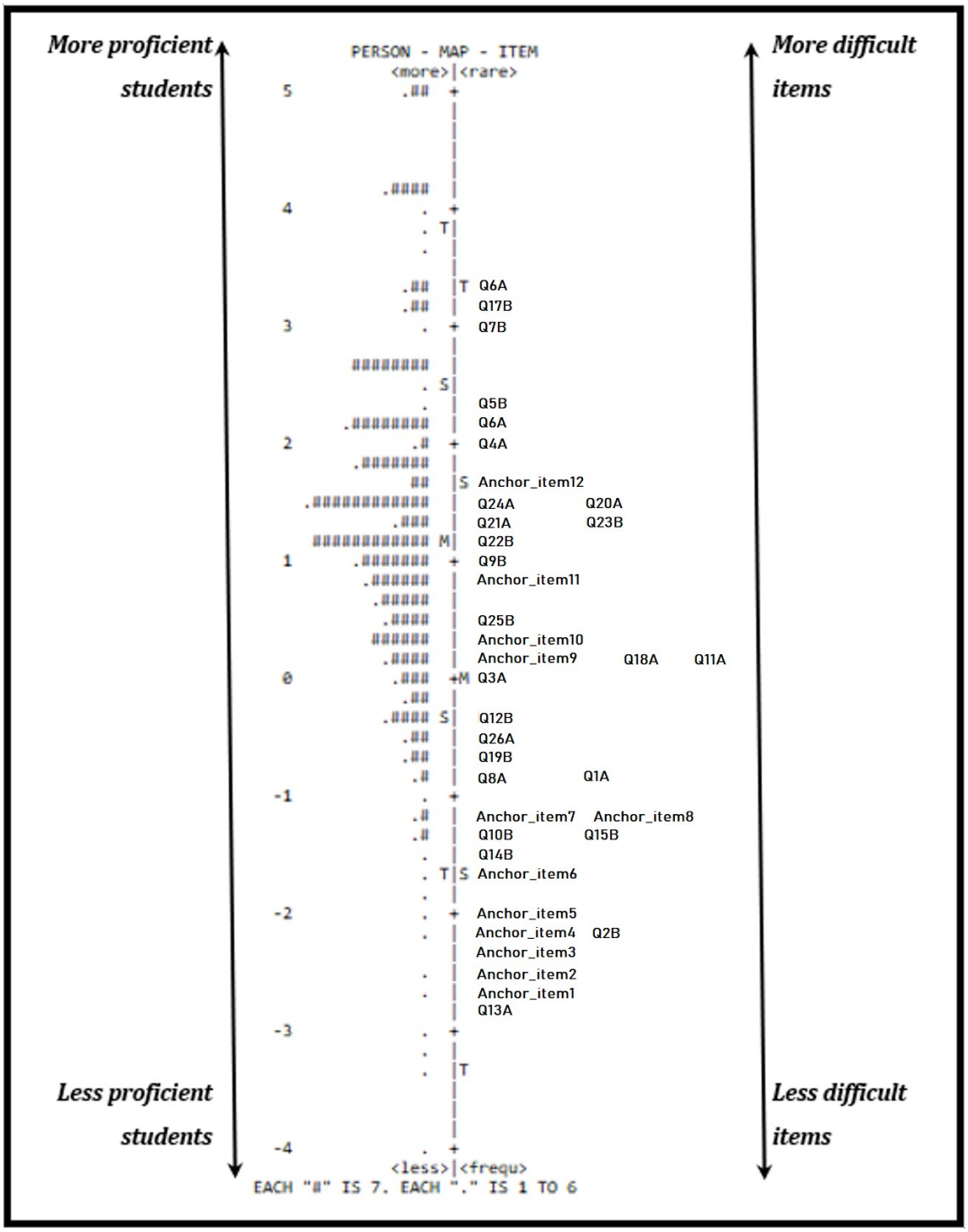

Figure 4. Person-Item map.