education sciences

MDPI

Article

Changing the Order of Factors Does Not Change the Product but Does Affect Students' Answers, Especially Girls' Answers

Clelia Cascella [1], Chiara Giberti [2,*] and Giorgio Bolondi [3]

[1] Cathie Marsh Institute for Social Research, University of Manchester, Manchester M13 9PL, UK; clelia.cascella@manchester.ac.uk
[2] Department of Human and Social Science, University of Bergamo, 24129 Bergamo, Italy
[3] Faculty of Education, Free University of Bolzano-Bozen, 39100 Bolzano, Italy; Giorgio.Bolondi@unibz.it
* Correspondence: chiara.giberti@unibg.it

Abstract: This study is aimed at exploring how different formulations of the same mathematical item may influence students' answers, and whether or not boys and girls are equally affected by differences in presentation. An experimental design was employed: the same stem-items (i.e., items with the same mathematical content and question intent) were formulated differently and administered to a probability sample of 1647 students (grade 8). All the achievement tests were anchored via a set of common items. Students' answers, equated and then analysed using the Rasch model, confirmed that different formulations affect students' performances and thus the psychometric functionality of items, with discernible differences according to gender. In particular, we explored students' sensitivity to the effect of a typical misconception about multiplication with decimal numbers (often called "multiplication makes bigger") and tested the hypothesis that girls are more prone than boys to be negatively affected by misconception.

Keywords: gender differences; mathematics achievement; item formulation; Rasch model; misconception; decimal numbers; multiplication

check for updates

Citation: Cascella, C.; Giberti, C.; Bolondi, G. Changing the Order of Factors Does Not Change the Product but Does Affect Students' Answers, Especially Girls' Answers. *Educ. Sci.* **2021**, *11*, 201. https://doi.org/10.3390/educsci11050201

Academic Editor: Andrew Lumpe

Received: 3 March 2021
Accepted: 16 April 2021
Published: 25 April 2021

1. Introduction

Differences in mathematical performance between boys and girls have received increasing attention over the years. Although the gap has narrowed over time, the issue is still topical since the differences continue to persist in many countries, as was reported by OECD-PISA (Organisation for Economic Co-operation and Development Programme for International Student Assessment) in 2015:

> "On average across OECD countries, boys outperform girls in mathematics by eight score points. Boys' advantage at the mean is statistically significant in 28 countries and economies" [1].

Most of the research studies carried out on this topic have used national or international large-scale assessment results and have operationalised gender differences as a reason behind the gap in mathematics test scores observed in relation to the entire test (e.g., [2–5]). Nevertheless, this perspective merely glances at gender differences, providing a snapshot of the gap between genders at some point or relating gender differences to other factors such as background and metacognitive aspects but failing to provide didactic information about the nature of these differences (differences that usually disadvantage girls more than boys), or explaining whether these differences are typically related to just some items or may concern all the test items. In this direction, part of the literature explores gender differences in relation to specific sub-domains of mathematical ability (for example, arguing that boys outperform girls in spatial ability and, more generally, in geometry items; e.g., [6,7]), other works at item level find a correlation between item difficulty and gender differences (e.g., [8,9]), and, finally, some studies examine the influence of item type in relation to gender (for instance, showing that boys outperform girls in multiple-choice items rather than constructed-response items, in which girls display better results; e.g., [10–13]). Less research has been carried out on possible relationships between task formulation and gender differences in relation to specific items, especially from a didactic perspective, considering the didactic milieu either as involved in the causes, or as an actor participating in the resolution. Let us state explicitly that throughout this paper, we use the term "gender" to indicate the result of the boys/girls classification used in the reports of the entities which have performed the studies, and the official registration of pupils used in Italian schools—it is a registry classification.

To explore these possible relationships, starting from a mathematics achievement test developed by the Italian National Institute for the Evaluation of Educational System (hereafter, INVALSI—Istituto Nazionale per la Valutazione del Sistema di Istruzione e Formazione) to measure students' ability in math at grade 8, we implemented an experimental plan. We prepared four booklets sharing some items, which are identical in all the booklets and compose the Core Test, while the remaining items are the same stem-items (i.e., items with the same mathematics content and the same question intent) formulated differently in each booklet. These variations were constructed to test specific hypotheses from mathematics education to explore if, and how, different formulations (mis)lead students' answers (and possibly problem-solving strategies), and subsequently to verify if and how this mechanism interplays with students' features (such as, for example, gender). In contrast to most of the current literature based on gender differences displayed over the entire test, as previously recommended, for example, by [14], we explored gender differences at item level, i.e., comparing the probability of encountering each item successfully by boys and girls matched on ability, via Rasch differential item functioning analysis (DIF). When we use the term "ability", related to the INVALSI test or to our experiment, we mean the latent trait measured globally by the test.

Our methodological strategy is two-fold: it tests didactic hypotheses about students' strategies and gathers some information about the effect of formulation and the activation of certain cognitive strategies over others, thus providing information about the relationship between formulation and item functionality from a psychometric point of view.

In this paper, we present an example of the analysis we carried out. Specifically, we explore the effect of a typical misconception about the multiplication of decimal numbers [15]. This misconception, often called "multiplication makes bigger" [16] (p. 37), emerges during the transition between natural numbers and decimal numbers: when operating with natural numbers, students see that the result of a multiplication is bigger than its factors, and suddenly, they begin to think that this property of multiplication is also true when they multiply rational numbers. This misconception leads students to also make mistakes in secondary school [15]. Previous studies have already proven that, in Italy, girls tend to conform their problem-solving strategies to didactic practices more than boys and could thus be more prone to the negative effect of misconceptions [17–21]. More generally, it is known that factors strictly related to the didactic choices of the teacher and of the school system have an impact on gender gap in mathematics. For instance, curriculum variables [22], teaching methods [23], different assessment practices [24], and factors related to achievement goals [25] are determinant in the emergence of gender differences in mathematical performance and that misconceptions in mathematics are related to intuitive models created during the didactic dialectic in the classroom [26]. The misconception we are exploring is evidently related to the model of multiplication as repeated addition, and hence to the order of the factors. In Italian, the first factor is the quantity to be multiplied, and the second factor is the "number of times", whilst in other languages (such as German), it is the opposite.

2. Theoretical Framework

2.1. Variation in the Formulation of a Mathematical Task

When students tackle a mathematics item, their answers are always influenced by the formulation of the item itself. Much research in mathematics education has studied the influence of the formulation specifically in mathematics word problems (e.g., [27]). Even minor changes in the formulation of a problem can affect students' answers. Many previous studies have already proven that the effects of variations in a task formulation are not simply related to linguistic formulation (in the case of word problems) but also to other variables such as data, context, and the operation involved. Nescher [28] proposed three categories of possible variations in a word problem: logical (operations involved, or lack or abundance of data), syntactic (number of words of the text, position of the question), and semantic (contextual relations and implicit suggestions). Duval [29] classified all these modifications as redactional variables, which influence students' cognitive and operative processes. Laborde [30] used this term to also include non-verbal changes, such as modification of figures or the position of the figures in relation to the text. A recent literature review on this issue [31] considered how linguistic variations as well as other kinds of changes influence students' responses and problem-solving strategies [32–34]. Daroczy [31] listed three main components that can alter the difficulty of a task, i.e., "(1) the linguistic complexity of the problem text itself, (2) the numerical complexity of the arithmetic problem, and (3) the relation between the linguistic and the numerical complexity of a problem" (p. 348). We may consider that even in a purely arithmetic task, the formulation may link it to intuitive models, and the usual contexts of use of the operations, which may affect its complexity.

For the purposes of this paper, we used the same question, "What is the result of 4×0.5?", previously administered by Sbaragli [15] (p. 124) but transformed into a multiple-choice item. In addition to the original form, we also administered another form with the same, but reversed, factors (0.5×4). We hypothesise that this change decreases the numerical complexity of the arithmetic problem because, in contrast to the original form, it suggests performing the multiplication following the intuitive model of repeated addition, with no conflict with the result, which is indeed higher than the first factor. This hypothesis is related to the fact that the students of our sample are Italian and in Italian, the first factor is the quantity to be multiplied, and the second factor is the "number of times". Moreover, this change in item formulation might increase item functionality.

In other words, our hypothesis is that the first formulation activates the misconception to a larger extent, and that this activation is stronger in girls.

2.2. Misconceptions and Decimal Numbers

During the early years of primary school, students learn natural numbers, their properties, and how to operate with them. The introduction of rational numbers is a complex phase and many difficulties emerge, primarily because rational numbers can be represented using different semiotic registers (e.g., fractions, decimal numbers, graphic representations). The literature shows that when students begin operating with decimal numbers, they have to overcome many obstacles [35,36].

The word misconception has been used with different meanings in the educational field [35], often as a synonym of "mistake" or "misunderstanding". Brousseau [37] linked misconceptions to the concept of "obstacle": during the formation of a mathematical concept, one idea that was useful earlier for solving problems can become an obstacle if students extend this idea to new problems where it is inappropriate. The mistake is due not to a lack of knowledge but to a previous knowledge that is incorrect in a more general context. When students study a new concept, they create an "intuitive model" of this concept [38] based on their primary experiences, but this model could be closer to the previous (more elementary) concept learned by the students in the past than to the complete mathematical concept, thus misleading students' problem-solving strategies.

When students learn natural numbers, they also learn properties, algorithms, and operations and, on this basis, create intuitive models of these concepts. Misconceptions related to this transition emerge, for example, when students compare decimal numbers and state that 0.12 is bigger than 0.2 just because 12 is bigger than 2 [15]: in this case, students compare the decimal part of the two numbers as if they were natural numbers. Moreover, they often do not consider 0.2 as 0.20 because, also in this case, they are influenced by the idea (correct in natural numbers but wrong with decimals) that adding zero at the end of the number is equal to multiplying it by 10.

The premature creation of intuitive models, indeed incomplete, and the persistence of these models leads students to make mistakes and generate "parasite" models [39]. In this paper, we adopt the following definition of misconception: a concept which is temporarily incorrect, awaiting re-elaboration in a more elaborated and critical cognitive system [39,40]. We focus on the misconception related to decimal numbers, according to which the result of a multiplication is always bigger than factors multiplied. This misconception has been widely studied in the literature and is usually called "multiplication makes bigger" [16,35,41]. It refers to the premature formation of a conceptual (intuitive) model of multiplication when students operate exclusively with natural numbers. When students learn multiplication, they use natural numbers, and then they observe that the product of two numbers (excluding 0 and 1) is always greater than its factors. This leads them to believe that the "rule" that "multiplication makes bigger" applies to both natural numbers and decimal numbers, although this is not actually true.

D'Amore and Sbaragli [35] interviewed students of different grades asking them "What is the result of 4 × 0.5?". The same question addressed to students attending primary, lower intermediate, and even secondary school was answered in a similar manner (i.e., 8) confirming that the mistake is due to the persistency of the misconception explained above. Our hypothesis is that we can correct the misconception by saying "multiplication makes the first factor bigger" (of course, in the Italian system).

3. Research Questions

Gender differences in mathematics test performance are explained in many studies by social and cultural factors (e.g., [3]) but also by metacognitive factors, such as a higher level of mathematics anxiety for girls and less self-confidence (e.g., [1]). These factors are also strictly related to the classroom environment, and previous studies based on INVALSI data showed that girls are more influenced by didactic practices, classroom routines, and the teacher–student relationship than boys, which makes them more prone to the (mis)leading effect of misconceptions and didactic contract [17–21,42].

A recent study argued that girls have more difficulties in solving items in which there is the influence of misconceptions on decimal numbers [17,19]. In particular, analysis of items that required comparison between decimal numbers showed that, when students work with decimal numbers with the same integer part (for example, 80.12 and 80.2), girls are more likely than boys to compare directly the decimal part of the two numbers and state that 80.12 is bigger than 80.2, probably considering that 12 is bigger than 2, rather than lower than 20.

Following these results, in this research, we study the previously described misconception, according to which the result of a multiplication is always bigger than its factors [15,16,41]. In order to explore this phenomenon, we compared two versions of the same stem-item; the first formulation was studied previously by Sbaragli in 2012 [15] via qualitative methods (Table 1).

Table 1. Different formulations of item D9.

Booklet F1 and F2	Booklet F3 and F4
D9. Which is the result of 4 x 0.5? Choose one of the following options. A. 8 B. 4 C. 2 D. 20	D9. Which is the result of 0.5 x 4? Choose one of the following options. A. 8 B. 4 C. 2 D. 20

Source: our elaboration.

In the second formulation, we simply reversed the order of the factors, in order to quantify the possible effect of the misconception described above from a gendered perspective. This variation was implemented in order to understand whether the misconception "multiplication makes bigger" is connected with both factors of the product (the result is bigger than both factors) or mostly to one of the two factors (i.e., the result is bigger than the first factor).

Our research questions are:

1. Does the misconception "multiplication always increases" have a different influence on boys and girls in terms of differential item functioning?
2. Does reversing factors (e.g., 4 × 0.5 in place of 0.5 × 4) have an impact on students' answers and item functionality?
3. Does this variation have a different effect on boys as compared with girls?

4. Materials and Methods

4.1. Data

A probability sample (2000 students attending grade 8), stratified by students' region of residence and socioeconomic (SES) background, was drawn from the entire list of schools located in Campania, Emilia-Romagna, Lazio, and Lombardy (four regions very representative of students' ability in the south, centre, and north of Italy, respectively, according to INVALSI national surveys—Figure 1). After data cleaning, the sample size equalled 1647 students, a number consistent with the Rasch equating design (roughly 400 students per form [43]).

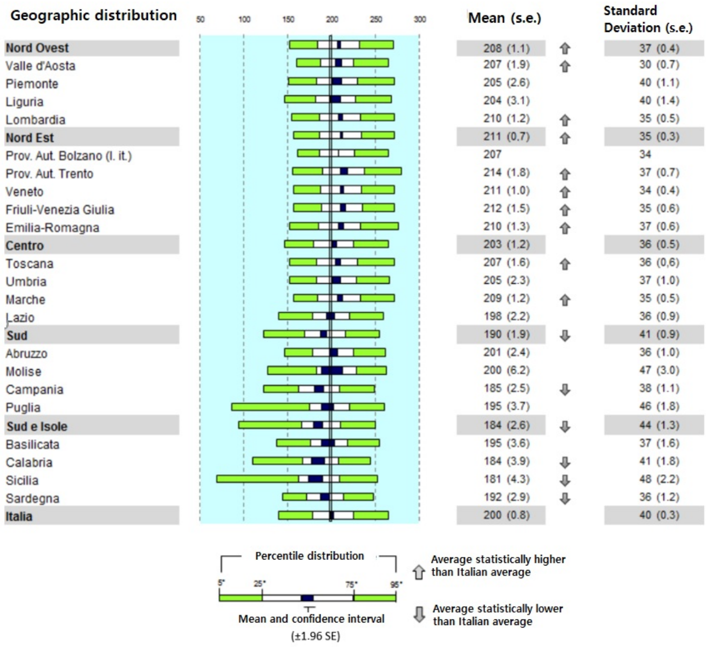

Figure 1. Scores distribution in mathematics INVALSI test at grade 8. Source: our adaptation from INVALSI National report 2017 [44].

To measure students' SES, the SC-index [45], based on the combination of highest parental education and professional status, was used. Individual SES data were aggregated at school level to measure overall school SES composition. The proportion of low-, medium-, and high-SES schools in our sample is similar to that in the annual INVALSI sample [44].

4.2. Materials

Using a mathematics achievement test developed by INVALSI as a starting point, three more achievement tests were developed. An experimental design was employed: all mathematics tests contained the same stem-items, i.e., items with the same mathematical content and the same question intent, but with a different formulation from one test to another. Item phrasing was modified by means of syntactic variations, different figures, the effect of mathematics, and/or real context.

In this paper, we analyse item D9, included in booklets F1 and F2 with the same formulation, and in F3 and F4 with altered phrasing. Variation was developed to explore the effect of misconception about multiplication with decimal numbers.

Table 2 shows the composition of each of the four booklets, the name of each item reporting the year of its inclusion in the INVALSI tests, and for the varied forms, we added "_original" or "_v" to specify in which booklet we included the original form or a varied form (different grey scale in the same row indicates different versions of a stem-item). In the first column, we indicate with "A" and "Anch" the items used to perform two anchoring strategies: (1) A first set of anchoring items was put at the beginning of the test in order to avoid the fatigue effect and offer an external anchoring strategy; these items are indicated by "A-". (2) A second set of anchoring items was included in the achievement test (in the same position across tests) and used as an internal anchor, indicated with "Anch-" (see Appendix A).

Table 2. Composition of the four booklets.

Item	Booklet 1	Booklet 2	Booklet 3	Booklet 4
A1a	D1a_PN2013	D1a_PN2013	D1a_PN2013	D1a_PN2013
A1b	D1b_PN2013	D1b_PN2013	D1b_PN2013	D1b_PN2013
A2	D18_PN2014	D18_PN2014	D18_PN2014	D18_PN2014
A3	D22_PN2013	D22_PN2013	D22_PN2013	D22_PN2013
A4	D10a_PN2012	D10a_PN2012	D10a_PN2012	D10a_PN2012
A5	D20_PN2010	D20_PN2010	D20_PN2010	D20_PN2010
A6	E18_PN2012	E18_PN2012	E18_PN2012	E18_PN2012
Anch_1	D7_PN2011	D7_PN2011	D7_PN2011	D7_PN2011
D1	D13_PN_2011_v4	D13_PN_2011_original	D13_PN_2011_v2	D13_PN_2011_v3
D2	D19_PN2011_original	D7_PN2011_v4	D7_PN2011_v3	D7_PN2011_v2
D3	E15_PN2012_original	E15_PN2012_v4	E15_PN2012_v3	E15_PN2012_v2
Anch_3	D18_PN2011	D18_PN2011	D18_PN2011	D18_PN2011
D5	D12_PN2011_original	D12_PN2011_v2	D12_PN2011_v3	D12_PN2011_v4
D6	D4_L052010_original	D4_L052010_v2	D4_L052010_v4	D4_L052010_v3
D7	D6_PN2011_original	D6_PN2011_v4	D6_PN2011_v3	D6_PN2011_v2
D8	D7b_L062013_v1	D7b_L062013_original	D7b_L062013_original	D7b_L062013_v1
Anch_7	D27_PN2013	D27_PN2013	D27_PN2013	D27_PN2013
D9	1CG_NEW_v1	1CG_NEW_v1	1CG_NEW_v2	1CG_NEW_v2
Anch_4	D17_PN2011	D17_PN2011	D17_PN2011	D17_PN2011
D10	1LG_NEW_v1	1LG_NEW_v2	1LG_NEW_v3	1LG_NEW_v4
Anch_8	D26_PN2015	D26_PN2015	D26_PN2015	D26_PN2015
Anch_5	D25_PN2011	D25_PN2011	D25_PN2011	D25_PN2011
D11	E6_PN2012_v3	E6_PN2012_v1	E6_PN2012_v2	E6_PN2012_v4
Anch_2	D9b_PN2011	D9b_PN2011	D9b_PN2011	D9b_PN2011
D12	D5_PN2011_original	D5_PN2011_v2	D5_PN2011_v4	D5_PN2011_v3
D13	E7_PN2012_v1	E7_PN2012_original	E7_PN2012_v4	E7_PN2012_v3
D14	D3_L062012_original	D3_L062012_v3	D3_L062012_v2	D3_L062012_v1
Anch_6	D22_PN2011	D22_PN2011	D22_PN2011	D22_PN2011
D15	3CG_NEW_v1	3CG_NEW_v1	3CG_NEW_v2	3CG_NEW_v2
D16	E16a_PN2012_original	E16a_PN2012_v2	E16a_PN2012_v1	E16a_PN2012_v3
D17	D8ab_PN2011_original	D8ab_PN2011_original	D8ab_PN2011_v3	D8ab_PN2011_v3

Items in white and labelled with "A" or "Anch" are anchored items; grey items labelled with "D" are items included with different formulations.

5. Analytic Strategy

The four mathematics achievement tests developed for the purposes of our research (named F1, F2, F3, and F4, respectively) were administered by means of a spiralling process (according to which different forms are administered to different students within each classroom) to randomly assign forms to students in the same classroom. Regarding the spiralling administration process: "When using this design, the difference between group level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms" [43] (p. 13) and thus is sufficient to render answers given by different subgroups of students comparable.

Nonetheless, to guarantee the comparability of answers provided by different students to the different versions of the same item (item D19), we scaled all students and all items from each achievement test along the same latent trait (i.e., mathematics ability) by anchoring our four mathematics achievement tests and then by equating them [43]. The process of equating is used in situations where scores earned on different forms need to be compared to each other. Within the Rasch framework [46], the process of equalising forms is used to construct a common scale and thus to put both students and items along the same latent trait, making them directly comparable.

In a recent study, Kopf, Zeileis, and Strobl [47] (p. 84) claimed that "The minimum (necessary but not sufficient) requirement for the construction of a common scale in the Rasch model is to place the same restriction on the item parameters in both groups [48]. The items included in the restriction are termed anchor items". Since the statistical power of anchoring increases with the length of a DIF-free (i.e., showing no differential functioning depending on students' features—[49]) anchor [50–52], we input two sets of anchor items: The first set of (eight) anchoring items was put at the beginning of the test in order to avoid a fatigue or learning effect, and then used for external anchoring; the other (eight) anchoring items were interspersed in between, through the test (in the same position across tests) and used as internal anchor. The first and the second set of anchor items were used as external and internal anchors in two separate calibration processes. Results from these anchoring strategies are consistent. Finally, both sets of items were used all together to perform a concurrent calibration to equate tests by using RUMM2030.

Having equalised the tests, a differential item functioning (DIF) analysis within the framework of the Rasch analysis was carried out to understand if, and how, misconception affects boys and girls differently.

The Rasch model is particularly suitable to pursue these aims as it grounds on three assumptions: (i) local independence (i.e., people's reactions to each item is independent from the reaction to all the other items); (ii) equal item discrimination (i.e., higher ability respondents are more likely to encounter each item successfully); and (iii) unidimensionality (i.e., a single common trait explains the item responses). To assess data-model fit, we preliminarily explored infit and outfit statistics, i.e., "mean-square fit statistics defined such that the model-specified uniform value of randomness is indicated by 1.0 [53]" (p. 9), with tolerable standard deviations around 0.20 [54]. Nonetheless, in line with previous studies (e.g., [55]) we took 1.3 as a value for infit and outfit mean squares that suggests cause of concern.

When these properties hold, Rasch parameters are invariance, i.e., they do not change across sub-group of students with the same level of ability. In contrast, violation of parameter invariance may be discovered by investigating the so-called differential item functioning (DIF; e.g., [56]). The DIF occurs when subjects matched on the same ability level have a different probability of encountering an item successfully. DIF refers to each single item and to item behaviour in a sub-group of students matched on ability and clustered by one personal student attribute (gender, in this study).

RUMM2030 compares the items' response function (IRF) that links the probability of a correct answer to student ability, for boys and girls separately. In fact, when a statistically significant DIF occurs, measurement invariance is violated, and thus, ''different item characteristics curves occur in subgroups" of students [47] (p. 83). In addition, we compared distractor response curves (DRC) drawn for boys and girls, separately. Distractor analysis is very informative because it provides a visual interpretation of response patterns for the set of distractors associated with each multiple-choice item. It allows examination of whether the differential selection of incorrect choices (distractors) attracts various groups in different ways (i.e., if any pattern is present in the proportion of responses across the different class intervals for each distractor against the IRF), thus identifying potential sources of construct-irrelevant variance. In addition, by comparing each distractor function, it was possible to examine whether variables other than a student's ability affect the content of only a single or all distractors. In addition, to assess statistically significance of gender differences, we reported on a factorial analysis of variance (ANOVA) on the class interval (factor 1) and person-level factors (factor 2).

6. Results

The following two sections report on DIF analysis by gender and on interpretation of output.

6.1. Differential Item Functioning by Gender

After having verified the goodness of data model fit (see results in the Appendix A), the item parameter estimate indicates the relative difficulty of item D9 and thus its location along the latent trait, a graded continuum where zero indicates a medium difficulty level, and thus, negative values indicate relatively easy items, whereas positive values indicate relatively more difficult items.

D9 is an easy item (δF1 = −1.094, SE 0.124; δF2 = −0.842, SE 0.118; δF3 = −1.668, SE 0.140; δF4 = −1.523, SE 0.132) and shows sharp differences (even though not always statistically significant—see Appendix A) between boys and girls matched on ability.

We provided a visual display of the set of observed means for each person level factor (i.e., for boys and girls) across each of the class intervals present in the item-trait test-of-fit specifications. Each level is plotted in relation to the item characteristic curve, i.e., the theoretical curve estimated by the Rasch model, according to which no factor other than students' intrinsic ability can affect the probability of encountering an item successfully. Finally, we reported the distractor plots drawn for boys and girls, separately, in order to explore their answer behaviour in relation to each answer option.

Boys outperform girls in relation to item formulation in booklet F1 (4 × 0.5) especially at the upper tail of the latent trait (i.e., among more talented students; Figures 2 and 3), although it shows both a non-significant interaction ($p = 0.739$) and a non-significant gender main effect ($p = 0.148$) ($\alpha = 0.05$—See Appendix A). Distractor analysis shows different students' approach to this item by gender—Option B is more attractive for low-ability boys than girls. Little difference can be found regarding options A and D.

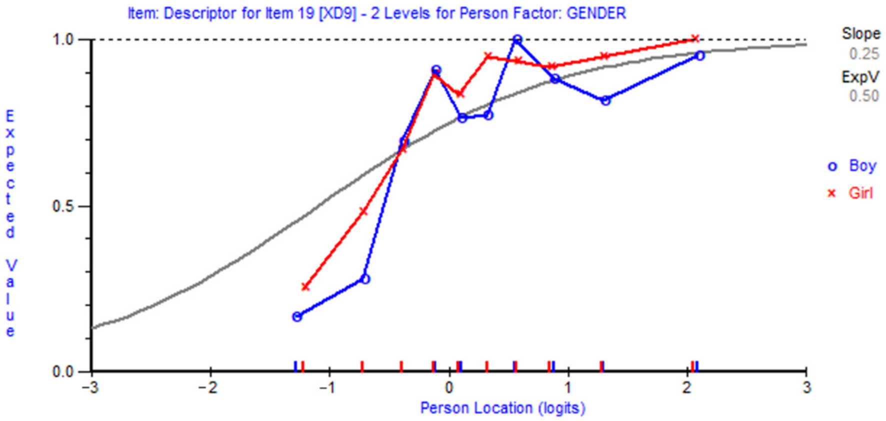

Figure 2. Item characteristic curve and DIF plot—item D9, booklet F1. Source: our elaboration.

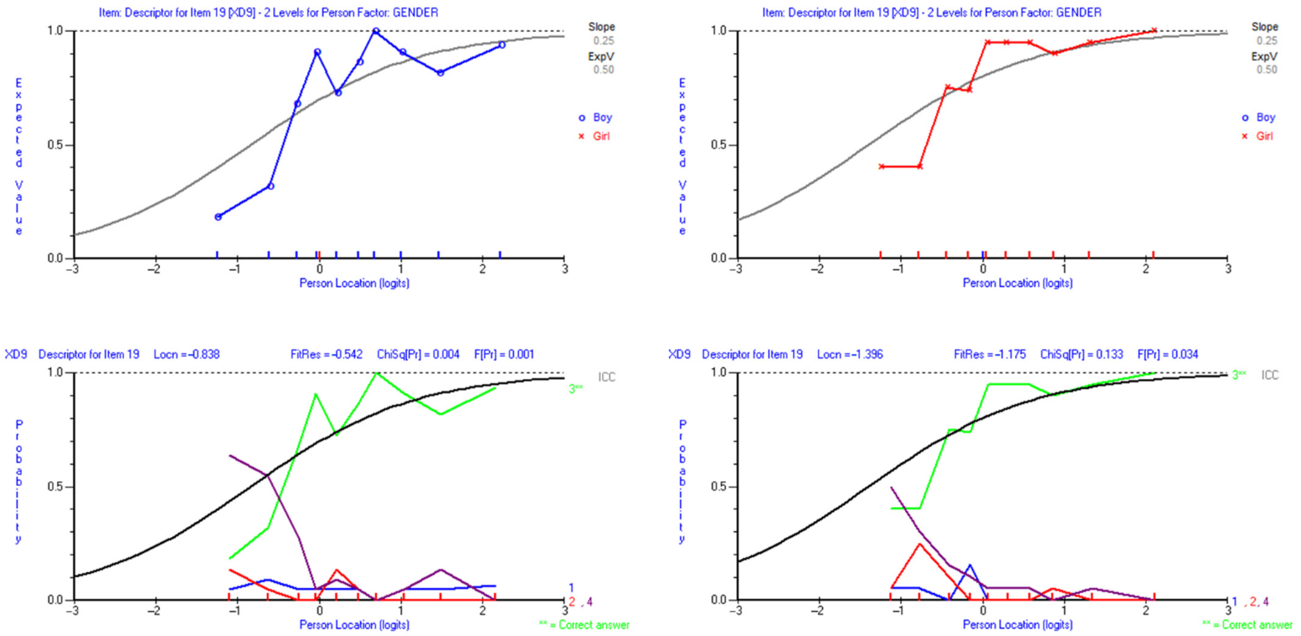

Figure 3. Distractor plot of boys (**bottom left**) and girls (**bottom right**)—item D9, booklet F1. Source: our elaboration. Note: The matrix has been split by gender. The figures above are the ICC plotter for boys (on the **left**) and girls (on the **right**). The figures below are the distractor plots for boys (on the **left**) and girls (on the **right**).

These differences are more notable in F2 (Figures 4 and 5), with a clear advantage of boys over girls, especially at the upper end of the latent trait, with a statistically significant main gender effect ($p = 0.001$) ($\alpha = 0.05$—See Appendix A). Nevertheless, response patterns for the set of distractors associated with D9 administered in the booklet F1 and D9 administered in the booklet F2 are naturally similar. The main difference relates to distractor D, which is much more attractive for girls (especially at the bottom of the ability distribution) than for boys. Moreover, high-ability boys are not attracted by any distractor.

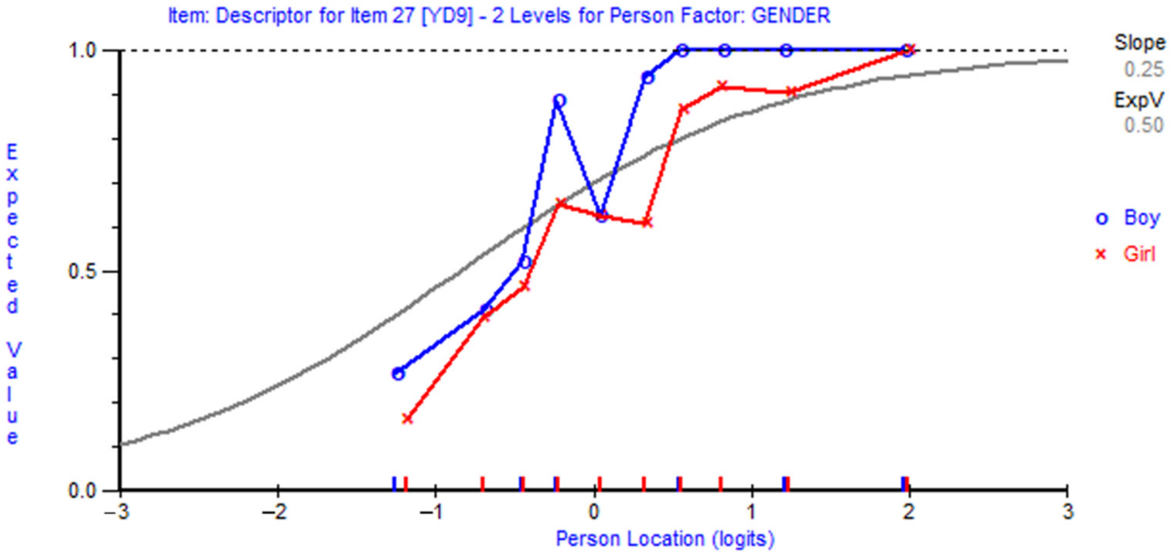

Figure 4. Item characteristic curve and DIF plot—item D9, booklet F2. Source: our elaboration.

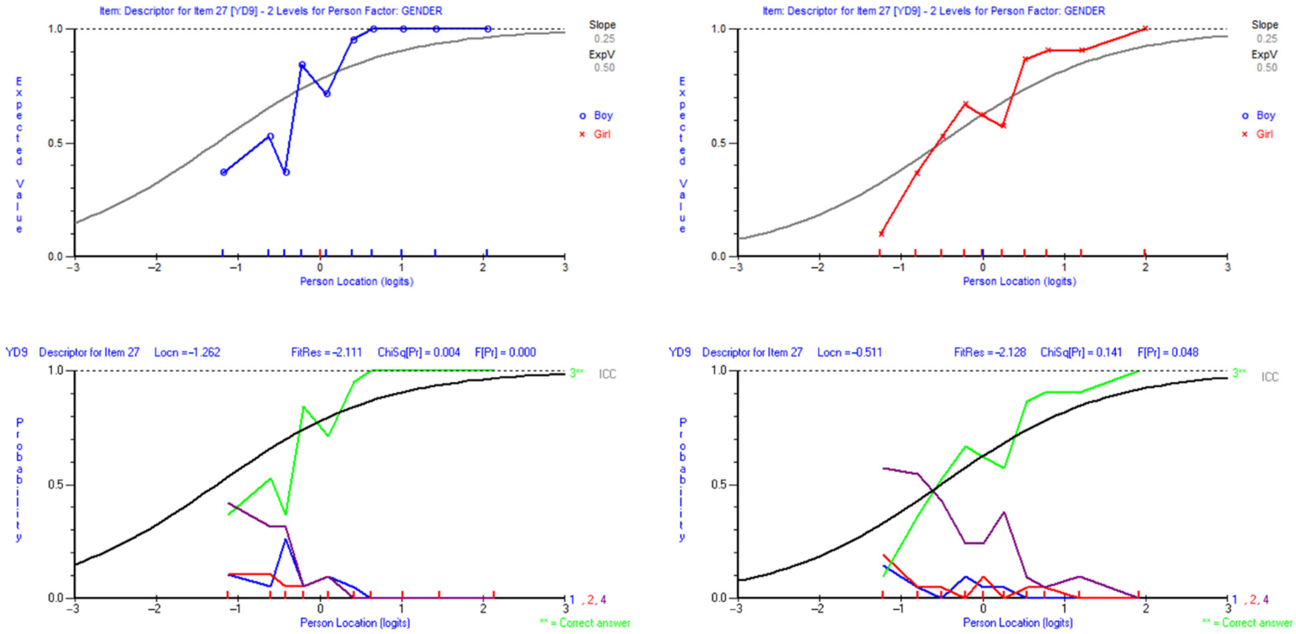

Figure 5. Distractor plot of boys (**bottom left**) and girls (**bottom right**)—item D9, booklet F2. Source: our elaboration. Note: The matrix has been split by gender. The figures above are the ICC plotter for boys (on the **left**) and girls (on the **right**). The figures below are the distractor plots for boys (on the **left**) and girls (on the **right**).

The analysis of answers to the item D9 administered in booklet F3 confirms an overall advantage of boys over girls (Figures 6 and 7), statistically significant in relation to the main gender effect in F3 (p = 0.027) and in relation to the interaction in F4 (p = 0.049) (α = 0.05—See Appendix A). The differences between boys and girls located at the bottom of the ability distribution are particularly interesting (Figure 8): from −1.5 to −0.5 logit, all differences are in favour of boys, as also partially confirmed by the analysis of F4. In both cases, distractor analysis shows interesting dissimilarities (Figure 9). The most interesting differences between boys and girls can be observed in F4. Option B is slightly more attractive for boys than for girls, while option D is much more attractive for low-ability girls.

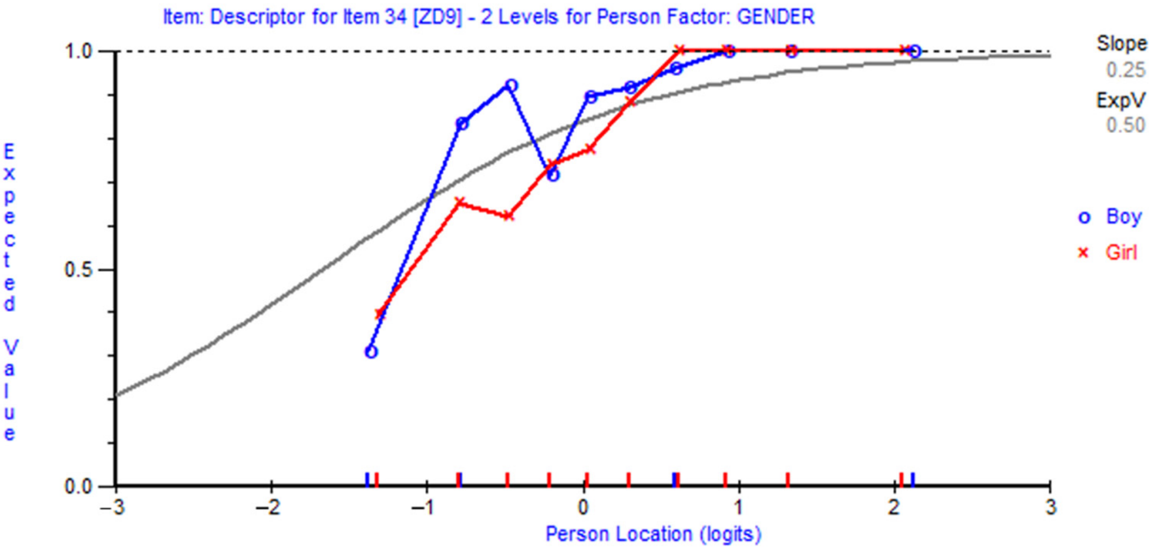

Figure 6. Item characteristic curve and DIF plot—item D9, booklet F3. Source: our elaboration.

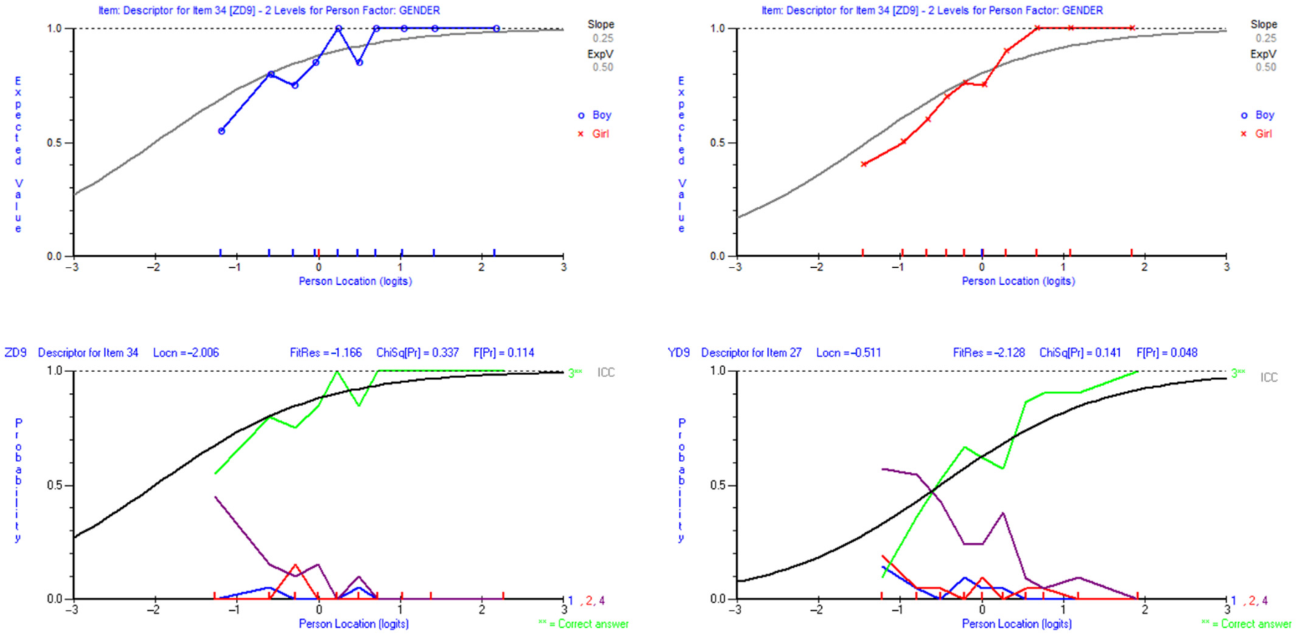

Figure 7. Distractor plot of boys (**bottom left**) and girls (**bottom right**)—item D9, booklet F3. Source: our elaboration. Note: The matrix has been split by gender. The figures above are the ICC plotter for boys (on the **left**) and girls (on the **right**). The figures below are the distractor plots for boys (on the **left**) and girls (on the **right**).

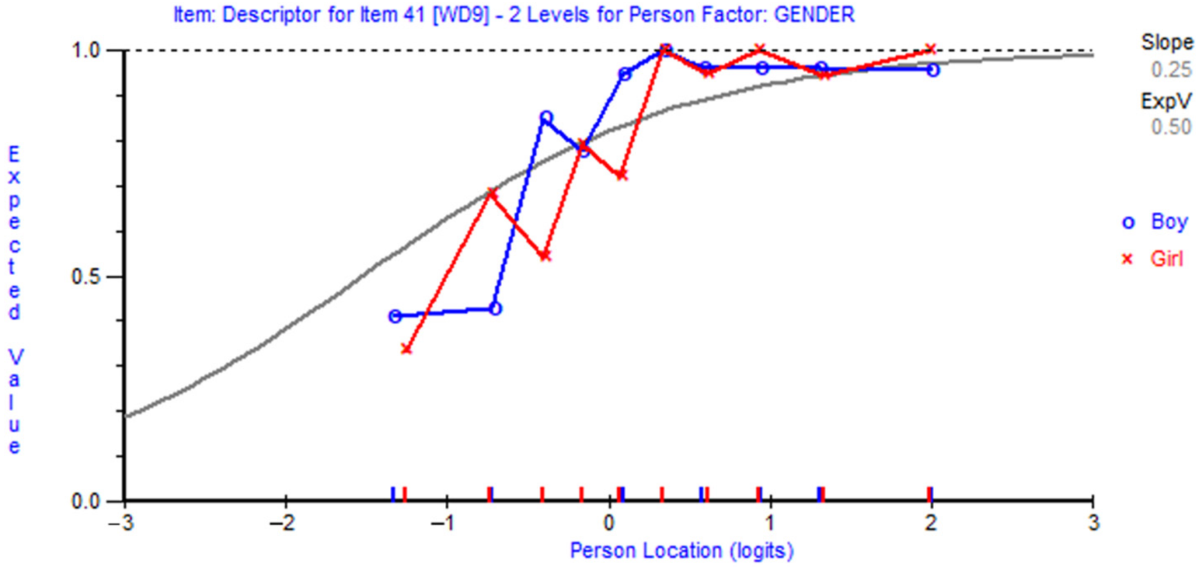

Figure 8. Item characteristic curve and DIF plot—item D9, booklet F4. Source: our elaboration.

6.2. The Interpretation of Empirical Results from a Didactic Point of View

The results reported above show interesting group differences. The graphical exploration and comparison of the ICCs and distractors as well as the comparison of item difficulty estimated for boys and girls provide some interesting elements that help to answer our research questions: misconception negatively affects the probability of encountering an item exploring students' ability in multiplying decimal numbers, and it affects girls more negatively than boys. The first item formulation (4 × 0.5) reveals, both in F1 and F2, an advantage for boys, especially at the bottom and top of the ability distribution.

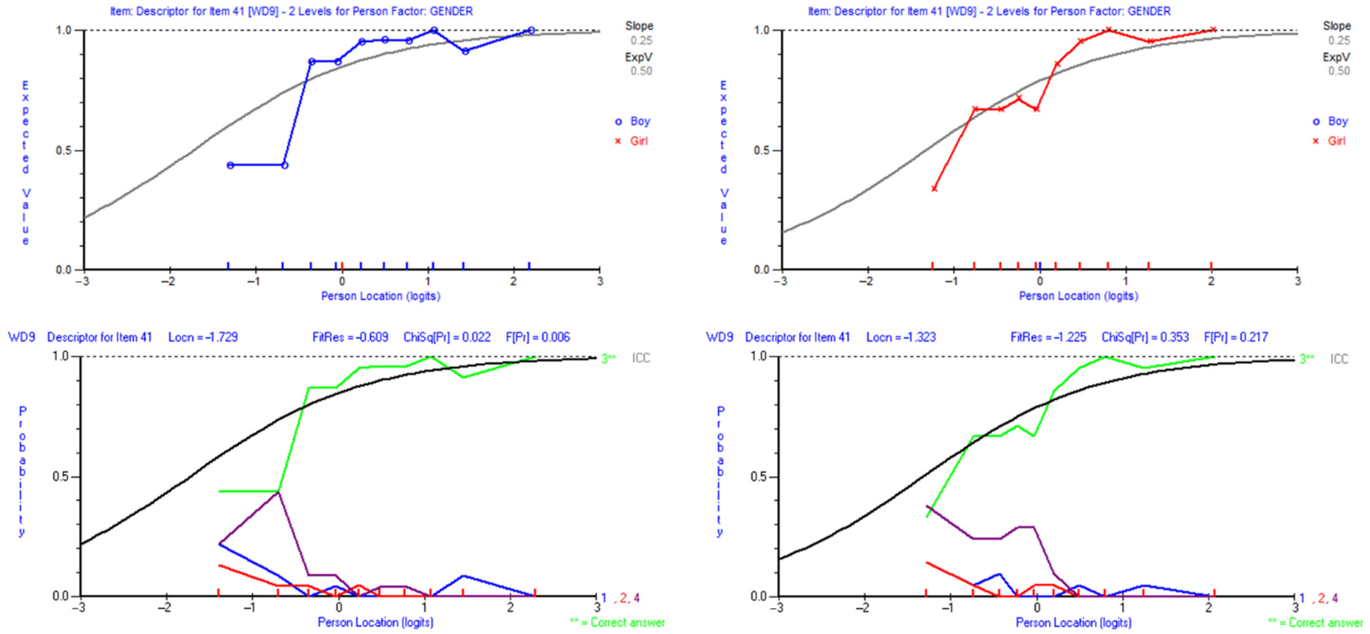

Figure 9. Distractor plot of boys (**bottom left**) and girls (**bottom right**)—item D9, booklet F4. Source: our elaboration. Note: The matrix has been split by gender. The figures above are the ICC plotter for boys (on the **left**) and girls (on the **right**). The figures below are the distractor plots for boys (on the **left**) and girls (on the **right**).

Comparison between the two versions of the task reveals that the order of the factors in multiplication has a strong influence on students' answers. In particular, if the multiplication posits the decimal number as the second factor (4 × 0.5), the task is more difficult than if the decimal number is presented first (0.5 × 4). This might be due to the fact that students are influenced by the intuitive model [38] of multiplication as a repeated sum, and in the second form, it is more immediate, for Italian students, to consider 0.5 × 4 = 0.5 + 0.5 + 0.5 + 0.5.

Therefore, the main finding is that the inversion of the two terms of a multiplication has a huge impact on students' behaviour, especially on girls: a stronger gender gap emerges in favour of boys in the first version (4 × 0.5) and in favour of girls in the second version (0.5 × 4) for the lower tail. This result is coherent with the fact that in the second version, it is easier, especially for struggling students, to tackle the task using the implicit model of multiplication as repeated addition, and this particularly helps girls of lower-ability levels.

7. Discussion and Conclusions

Limitation of the Present Study

The DIF analysis revealed gender differences, not always statistically significant. This could be a limitation of the present study because results presented in this paper cannot be inferred to the entire student population. Nonetheless, it is worth noting that the Rasch model is based on the assumption that the probability of encountering an item successfully is related to students' relative ability, i.e., their ability compared with item difficulty, and that no other variables (e.g., students' individual features) can affect it. Therefore, even though a moderate item misfit does not need to necessarily be interpreted as a limitation (of the test or even of the choice of the model), but as a potential source of information (as recently argued in [57]), test items are constructed by INVALSI to be DIF-free. Similarly, the materials we developed for the purposes of the present research were constructed to be DIF-free with just a few exceptions aimed at testing specific hypotheses about how gender interplays with item characteristics. Nonetheless, only three items were constructed to explore gender differences (D9, D15, and D16, aimed at exploring misconceptions or the effect of the item's context—i.e., real or mathematical—on students' solving strategy). The absence of a statistically significant DIF is thus an unavoidable and inherent consequence of the tests' construction process.

Moreover, even though our analysis revealed some differences in students' answers to the item in F1 and F2, and to the item in F3 and in F4, it is worth noting that results between F1 and F2 are consistent, as are those between F3 and F4, thus supporting our results' interpretations about the diverse effect of the misconception analysed in this paper on girls' and boys' answers.

Results presented in this paper showed that traditional psychometric tools, and in particular the graphical inspection of the ICCs and of distractor plots, are extremely valuable in exploring in-group differences, since all the graphs compare students matched on ability. Moreover, in this research, such graphs were constructed after having equated mathematics achievement tests, thus making students' answers directly comparable. Working within the framework of the Rasch analysis is an added value of the present study: the equating strategy performed here guarantees the comparability of students' answers across mathematics achievement tests and across sub-groups of students (whichever way they are defined), thus offering a methodological approach that can be used also to pursue other research goals.

Our analyses showed a different effect of a specific misconception (related to multiplication with decimal numbers) on boys' and girls' answering behaviour. The misconception investigated here was already studied from a qualitative point of view by D'Amore and Sbaragli [35]. Consistently with Sbaragli [15], our results showed that girls' difficulty in multiplying decimal numbers is due to the misconception, as also confirmed by distractor D, which is strictly related to the misconception and is more attractive for girls than for boys.

The inversion of multiplication factors misleads students' answers, with a stronger influence on girls than on boys. Moreover, compared to previous studies about students' misconceptions in multiplying decimal numbers, the use of the Rasch model adds some advantages to the investigation of this topic. Firstly, if DIF is detected, the results can be interpreted in terms of which items are easier or harder to solve for which group [47]. This offers interesting elements to enrich the debate from a didactic point of view: previous studies carried out in Italy have shown that girls are more influenced by didactic practices, classroom routines, and the teacher–student relationship than boys, and that this makes them more prone to the (mis)leading effect of misconceptions and didactic contract [16,38]. Moreover, such strong differences between boys and girls at grade 8 in Italy are quite unusual: as systematically reported by INVALSI in its national annual reports, gender differences increase over time, from primary to secondary school, but at grade 8, they tend to be close to zero (e.g., [44]). Understanding such a result deserves much more investigation that is beyond the scope of this study.

Results presented in this paper help us to explain why and how the exploration of gender gap at item level, rather than across the entire test, can contribute further information to the current debate about gender differences. In this direction, for example, Leder and Lubiensky [14] (p. 35) stated that:

> Item-level analyses can pinpoint the mathematics that students do and do not know, including which problems most students can and cannot solve, and which problems have the largest disparities between groups. This information can inform both textbook writers and teachers, as they strive to address curricular areas in need of additional attention. Hence, it is important for item-level analyses to be systematically conducted and reported.

In this paper, we combined traditional psychometrical tools with the theoretical lens of mathematics education to test specific hypotheses about students' problem-solving strategies. This comparison, based on a large probability sample consisting of 1647 students attending grade 8, was made on the analysis of students' answers to four anchored mathematics tests developed for the specific purposes of the present study. A common-item non-equivalent group design was employed to collect data, and all forms were equalised to enable comparable answers from the different subgroups of students: "When using this design, the difference between group level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms" [39] (p. 13).

The combination of traditional psychometric tools with the theoretical lens of mathematics education, an unprecedented strategy for the exploration of gender differences, adds real value to the current debate about gender differences because it provides critical information about boys' and girls' performances and hence suggests research paths about their problem-solving strategies. Gender differences emerge on specific mathematics content, and these results are consistent with the current literature on gender differences in mathematics: many studies highlight that differences between boys and girls can be explained by a different use of learning and problem-solving strategies rather than differences in cognitive abilities. If we consider problem-solving activities in mathematics, for instance, girls more frequently use routine procedures and well-known algorithms, while boys are more inclined to try new methods and non-conventional approaches [58–60]. The analysis of gender difference in items related to specific difficulties and constructs already studied in mathematics education research could be fruitful also for teachers. The more we investigate and understand these differences, the more teachers will have opportunities to intervene with specific didactical activities. In particular, regarding misconception, teachers must be aware of avoidable and unavoidable misconception [15]. The first ones are linked to didactical practices and teachers' choices; the second ones are unavoidable because they are not due to didactical transposition but are temporary and not exhaustive ideas due to the necessary gradual introduction of new mathematical knowledge. In this paper, we compared two versions of the same item with the purpose of analysing a specific misconception concerning multiplication with decimal numbers. Misconceptions related to decimal numbers are considered unavoidable misconceptions: they arise from the fact that students learn mathematical operation in the field of natural numbers. Teachers must be conscious of students' difficulties in the transition between natural and rational numbers: they need to ensure that ideas related to mathematics operations in natural numbers do not become "parasite" models [15] when students have to face the same operation with rational numbers. In particular, this study suggests to teachers to pay special attention to girls because they are more influenced by these misconceptions. In our task, we observe that differential item functioning is related to misconceptions and intuitive models of multiplications used by students, but the influence of these factors is different for boys and girls. This is further confirmed by variation in item functionality due to variation in item formulation: 0.5×4 favours lower-ability girls by offering an "easier" formulation which activates a routine procedure (intuitive model of multiplication as repeated addition).

This paper gives a contribution in the direction indicated by [61]: a theoretically driven interpretation of macrophenomena highlighted quantitatively by Large-Scale Assessments may help in clarifying solid findings in Mathematical Education.

Author Contributions: Conceptualization, C.G. and G.B.; Data curation, C.C.; Funding acquisition, C.C.; Methodology, C.C.; Software, C.C.; Supervision, G.B.; Writing–original draft, C.C., C.G. and G.B.; Writing–review & editing, C.C., C.G. and G.B. All the authors contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and according to the Ethical code of the Free University of Bozen.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We thank all the students and teachers involved in the experimentation and INVALSI for the support in the implementation of this research. Nonetheless, the views expressed in this work are attributable exclusively to the authors, and thus, in citing the work, it is not appropriate to attribute the arguments expressed herein to INVALSI.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The table below reports on results from the analysis of variance (described in the Methodological Section), carried out using RUMM2030 (the significance level used was $\alpha = 0.05$).

Table A1. Analysis of variance for item D9 in booklet F1.

Source	S.S	DF	MS	F-Ratio	Prob
Between	33.432	19	1.760		
ANOVA-fit[c-int]	27.755	9	3.084	4.400614	0.000018
DIF [gender]	1.475	1	1.475	2.104466	0.147681
Gender-by-Cinf	4.203	9	0.467	0.666371	0.739415
Total item DIF	5.678	10	0.568	0.810180	0.619006
Total Misfit	33.432	19	1.760	2.510912	0.000497
Within	269.097	384	0.701		
Total	302.529	403	0.751		

Source: our elaboration.

Table A2. Analysis of variance for item D9 in booklet F2.

Source	S.S	DF	MS	F-Ratio	Prob
Between	40.188	19	2.115		
ANOVA-fit[c-int]	28.694	9	3.188	4.712766	0.000005
DIF [gender]	7.656	1	7.656	11.316170	0.000854
Gender-by-Cinf	3.838	9	0.426	0.630367	0.771172
Total item DIF	11.494	10	1.149	1.698947	0.079041
Total Misfit	40.188	19	2.115	3.126546	0.000016
Within	257.074	380	0.677		
Total	297.262	399	0.745		

Source: our elaboration.

Table A3. Analysis of variance for item D9 in booklet F3.

Source	S.S	DF	MS	F-Ratio	Prob
Between	25.904	19	1.363		
ANOVA-fit[c-int]	18.558	9	2.062	3.040102	0.001579
DIF [gender]	3.343	1	3.343	4.928163	0.026998
Gender-by-Cinf	4.003	9	0.445	0.655813	0.748833
Total item DIF	7.346	10	0.735	1.083048	0.374000
Total Misfit	25.904	19	1.363	2.010073	0.007471
Within	261.138	385	0.678		
Total	287.042	404	0.711		

Source: our elaboration.

Table A4. Analysis of variance for item D9 in booklet F4.

Source	S.S	DF	MS	F-Ratio	Prob
Between	34.804	19	1.832		
ANOVA-fit[c-int]	20.644	9	2.294	3.031639	0.001595
DIF [gender]	1.164	1	1.164	1.538673	0.215518
Gender-by-Cinf	12.996	9	1.444	1.908551	0.049147
Total item DIF	14.160	10	1.416	1.871563	0.047361
Total Misfit	34.804	19	1.832	2.421072	0.000802
Within	314.751	416	0.757	0.047361	
Total	349.555	435	0.804	0.000802	

Source: our elaboration.

References

1. OECD. *PISA 2015 Results (Volume I): Excellence and Equity in Education*; PISA, OECD Publishing: Paris, France, 2016.
2. Contini, D.; Di Tommaso, M.L.; Mendolia, S. The gender gap in mathematics achievement: Evidence from Italian data. *Econ. Educ. Rev.* **2017**, *58*, 32–42. [CrossRef]
3. Guiso, L.; Monte, F.; Sapienza, P.; Zingales, L. Culture, gender, and math. *Science* **2008**, *320*, 1164. [CrossRef]
4. Guo, J.; Marsh, H.W.; Parker, P.D.; Morin, A.J.; Yeung, A.S. Expectancy-value in mathematics, gender and socio-economic background as predictors of achievement and aspirations: A multi-cohort study. *Learn. Individ. Differ.* **2015**, *37*, 161–168. [CrossRef]
5. Rodríguez-Planas, N.; Nollenberger, N. Let the girls learn! It is not only about math ... it's about gender social norms. *Econ. Educ. Rev.* **2018**, *62*, 230–253. [CrossRef]
6. Harris, A.M.; Carlton, S.T. Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Appl. Meas. Educ.* **1993**, *6*, 137–151. [CrossRef]
7. Lawton, C.A.; Hatcher, D.W. Gender differences in integration of images in visuospatial memory. *Sex Roles* **2005**, *53*, 717–725. [CrossRef]
8. Bielinski, J.; Davison, M.L. A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *J. Educ. Meas.* **2001**, *38*, 51–77. [CrossRef]
9. Penner, A.M. International gender by item difficulty interactions in mathematics and science achievement tests. *J. Educ. Psychol.* **2003**, *95*, 650–655. [CrossRef]
10. Anderson, J. Gender-related differences on open and closed assessment tasks. *Int. J. Math. Educ. Sci. Technol.* **2002**, *33*, 495–503. [CrossRef]
11. Bolger, N.; Kellaghan, T. Method of Measurement and Gender Differences in Scholastic Achievement. *J. Educ. Meas.* **1990**, *27*, 165–174. [CrossRef]
12. DeMars, C.E. Test Stakes and Item Format Interactions. *Appl. Meas. Educ.* **2000**, *13*, 55–77. [CrossRef]
13. Pomplun, M.; Capps, L. Gender differences for constructed-response mathematics items. *Educ. Psychol. Meas.* **1999**, *59*, 597–614. [CrossRef]
14. Leder, G.; Lubienski, S. Large-Scale Test Data: Making the Invisible Visible. In *Diversity in Math Education*; Springer: Cham, Switzerland, 2015; pp. 17–40.
15. Sbaragli, S. Il ruolo delle misconcezioni nella didattica della matematica. In *I Quaderni della Didattica. Metodi e Strumenti per l'insegnamento e L'apprendimento Della Matematica*; Bolondi, G., Fandiño Pinilla, M.I., Eds.; EDISES: Napoli, Italy, 2012; pp. 121–139.
16. Greer, B. Nonconservation of multiplication and division involving decimals. *J. Res. Math Educ.* **1987**, *18*, 37–45. [CrossRef]
17. Bolondi, G.; Cascella, C.; Giberti, C. Highlights on gender gap from Italian standardized assessment in mathematics. In *Diversity in Mathematics Education*; Novotnà, J., Moravà, H., Eds.; Universita Karlova Press: Prague, Czech Republic, 2017.
18. Bolondi, G.; Ferretti, F.; Giberti, C. Didactic Contract as a Key to Interpreting Gender Differences in Maths. *Ecps Educ. Cult. Psychol. Stud.* **2018**, *18*, 415–435. [CrossRef]
19. Cascella, C.; Giberti, C.; Bolondi, G. A Differential Item functioning analysis to explore gender gap in math tasks. Studies in Educational Evaluation. *Stud. Educ. Eval.* **2020**, *64*, 100819. [CrossRef]
20. Ferretti, F.; Giberti, C.; Lemmo, A. The Didactic Contract to Interpret Some Statistical Evidence in Mathematics Standardized Assessment Tests. *Eurasia J. Math. Sci. Technol. Educ.* **2018**, *14*, 2895–2906. [CrossRef]
21. Giberti, C. Differenze di genere e misconcezioni nell'operare con le percentuali: Evidenze dalle prove INVALSI. *CADMO* **2019**, *2*, 97–114. [CrossRef]
22. Leder, G.C. Mathematics and gender: Changing perspectives. In *Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics*; Grouws, D.A., Ed.; Macmillan Publishing Co, Inc.: New York, NY, USA, 1992; pp. 597–622.
23. Boaler, J. Learning from Teaching: Exploring the Relationship between Reform Curriculum and Equity. *J. Res. Math. Educ.* **2002**, *33*, 239. [CrossRef]
24. Leder, G.C.; Forgasz, H.J. Mathematics education: New perspectives on gender. *ZDM* **2008**, *40*, 513–518. [CrossRef]

25. Wirthwein, L.; Sparfeldt, J.R.; Heyder, A.; Buch, S.R.; Rost, D.H.; Steinmayr, R. Sex differences in achievement goals: Do school subjects matter? *Eur. J. Psychol. Educ.* **2019**, *35*, 1–25. [CrossRef]
26. Sbaragli, S. Le misconcezioni in aula. In *Dal Pensare Delle Scuole: Riforme*; Boselli, G., Seganti, M., Eds.; Armando Editore: Roma, Italy, 2006; pp. 130–139.
27. Verschaffel, L.; Greer, B.; De Corte, E. Making sense of word problems. *Educ. Stud. Math* **2000**, *42*, 211–213.
28. Nesher, P. Levels of description in the analysis of addition and subtraction word problems. In *Addition and Subtraction: A cognitive Perspective*; Carpenter, T.P., Moser, J.M., Romberg, T.A., Eds.; Routledge: Oxford, UK, 1982; pp. 25–38.
29. Duval, R. Interaction des différents niveaux de représentation dans la compréhension de textes. *Annal. Didact. Sci. Cognit.* **1991**, *4*, 136–193.
30. Laborde, C. Occorre apprendere a leggere e scrivere in matematica. *La Mat. E La Sua Didatt.* **1995**, *9*, 121–135.
31. Daroczy, G.; Wolska, M.; Meurers, W.D.; Nuerk, H.-C. Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Front. Psychol.* **2015**, *6*, 348. [CrossRef] [PubMed]
32. D'Amore, B. Lingua, matematica e didattica. *Mat. Didatt.* **2000**, *1*, 28–47.
33. De Corte, E.; Verschaffel, L.; Van Coillie, V. Influence of number size, problem structure, and response mode on chil-dren's solutions of multiplication word problems. *J. Math. Behav.* **1988**, *7*, 197–216.
34. Thevenot, C.; Devidal, M.; Barrouillet, P.; Fayol, M. Why does placing the question before an arithmetic word problem improve performance? A situation model account. *Q. J. Exp. Psychol.* **2007**, *60*, 43–56. [CrossRef]
35. D'Amore, B.; Sbaragli, S. Analisi semantica e didattica dell'idea di "misconcezione". *Mat. Didatt.* **2005**, *2*, 139–163.
36. Steinle, V.; Stacey, K. The incidence of misconceptions of decimal notation amongst students in Grades 5 to 10. In *Teaching Math in New Times. Proceedings of the 21st Annual Conference of the Math Education Research Group of Australasia*; Kanes, C., Goos, M., Warren, E., Eds.; MERGA: Gold Coast, Australia, 1998; Volume 2, pp. 548–555.
37. Brousseau, G. Les obstacles épistémologiques et les problèmes en mathématiques. *Rech. Didact. Mat.* **1983**, *4*, 165–198.
38. Fischbein, E.; Deri, M.; Nello, M.S.; Marino, M.S. The role of implicit models in solving verbal problems in multipli-cation and division. *J. Res. Math Educ.* **1985**, *16*, 3–17. [CrossRef]
39. D'Amore, B. *Elementi di Didattica della Matematica*; Pitagora: Bologna, Italy, 1999.
40. Maier, H.; Bauma, P.G. Book Review: Elementi di Didattica della Matematica. *ZDM* **2001**, *33*, 103.
41. Hart, K.M.; Brown, M.L.; Kuchemann, D.E.; Kerslake, D.; Ruddock, G.; McCartney, M. *Children's Understanding of Math: 11–16*; John Murray: London, UK, 1981.
42. Ferretti, F.; Giberti, C. The Properties of Powers: Didactic Contract and Gender Gap. *Int. J. Sci. Math. Educ.* **2020**, 1–19. [CrossRef]
43. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2004.
44. INVALSI. Rilevazione Nazionale Degli Apprendimenti 2016–2017. 2017. Available online: http://www.invalsi.it/invalsi/doc_eventi/2017/Rapporto_Prove_INVALSI_2017.pdf (accessed on 22 February 2021).
45. Cascella, C. How much does 'home possession' affect educational attainment? Empirical evidences towards a simpler bi-dimensional SES index. In Proceedings of the ICERI 2019 Conference, Seville, Spain, 11–13 November 2019; pp. 5809–5814, ISBN 978-84-09-14755-7.
46. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Denmarks Paedagogiske Institut: Copenhagen, Denmark, 1960.
47. Kopf, J.; Zeileis, A.; Strobl, C. A framework for anchor methods and an iterative forward approach for DIF detection. *Appl. Psychol. Meas.* **2015**, *39*, 83–103. [CrossRef] [PubMed]
48. Glas, C.A.W.; Verhelst, N.D. Testing the Rasch Model. In *Rasch Models*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 1995; pp. 69–95.
49. Osterlind, S.J.; Everson, H.T. *Differential Item Functioning*; Sage Publications: Thousand Oaks, CA, USA, 2009; Volume 161.
50. Shih, C.L.; Wang, W.C. Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Appl. Psychol. Meas.* **2009**, *33*, 184–199. [CrossRef]
51. Wang, W.C. Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *J. Exp. Educ.* **2004**, *72*, 221–261. [CrossRef]
52. Woods, C.M. Empirical selection of anchors for tests of differential item functioning. *Appl. Psychol. Meas.* **2009**, *33*, 42–57. [CrossRef]
53. Wright, B.; Panchapakesan, N. A procedure for sample-free item analysis. *Educ. Psychol. Meas.* **1969**, *29*, 23–48. [CrossRef]
54. Engelhard, G., Jr. *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*; Routledge: London, UK, 2013.
55. Cascella, C.; Pampaka, M. Attitudes Towards Gender Roles in Family: A Rasch-based Validation Study. *J. Appl. Meas.* **2020**, *21*, 2020.
56. Ackerman, T.A. A didactic explanation of item bias, item impact, and item validity from a multidimensional perspec-tive. *J. Educ. Meas.* **1992**, *29*, 67–91. [CrossRef]
57. Bolondi, G.; Cascella, C. A mixed approach to interpret Large-Scale assessment psychometric results of the learning of Mathematics. *Mat. Didatt.* **2020**, *28*, 1–21.

58. Bell, K.N.; Norwood, K. Gender equity intersects with math and technology: Problem-solving education for changing times. In *Gender in the Classroom*; Mahwah, N.J., Sadker, D., Silber, E.S., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2007; pp. 225–258.
59. Fennema, E.; Carpenter, T.P.; Jacobs, V.R.; Franke, M.L.; Levi, L.W. New perspectives on gender differences in math: A reprise. *Educ. Res.* **1998**, *27*, 19–21. [CrossRef]
60. Gallagher, A.M.; De Lisi, R.; Holst, P.C.; Lisi, A.V.M.-D.; Morely, M.; Cahalan, C. Gender Differences in Advanced Mathematical Problem Solving. *J. Exp. Child Psychol.* **2000**, *75*, 165–190. [CrossRef] [PubMed]
61. Bolondi, G.; Ferretti, F. Quantifying Solid Findings in Mathematics Education: Loss of Meaning for Algebraic Symbols. *Int. J. Inn. Sci. Math. Edu.* **2021**, *29*, 1–15.