How much item formulations affect the probability of a correct answer? An experimental study.

Giorgio Bolondi[1], Clelia Cascella[2], Chiara Giberti[1]

[1]*Free University of Bolzano-Bozen,* [2]*University of Manchester*

Different words, numeric values or semiotic registers, figures, graphs or tables, namely elements of item formulation, affect students' probability of encountering an item correctly. An experimental study was carried out to compare different formulations and their effect for validating a comparative technique: students' answers were equated within the framework of the Rasch analysis to make them directly comparable on the same scale and then analyzed and compared to each other. Four anchored math achievement tests were administered to a sample of 1647 students attending grade 8. In particular, we used different formulations of the same items to explore students' misconception about the relationship between perimeter and area. Results confirmed that item formulation channels students' solving strategy and thus modifies the probability of a correct answer more than item content. This quantitative technique combined in a mixed-method approach with qualitative analysis may provide useful didactical issues.

Keywords: Misconception; Large-scale assessment; Rasch model

Introduction

The purpose of this paper is to present a quantitative methodology for approaching the problem of how the formulation of an item affects student processes and student performances. This quantitative approach can be used in a mixed methodology, since it may provide evidences about new phenomena, which either can be investigated via a qualitative methodology or can give further insights concerning known results coming from previous research.

Our methodology is based on an anchoring technique and the Rasch Model, which is presented in the methodology chapter together with the experimental plan. In particular, we studied seventeen mathematical items, each varied in its formulation in two or four forms. As a theoretical framework for the problem of formulation in word problem we referred to Daroczy et al. (2015) and to the discussion carried on in Bolondi, Branchetti and Giberti (2018).

In the discussion chapter we present the results of our experiment in a specific case: an item where a well studied misconception about area and perimeter is involved. Our technique shows that the percentage of students influenced by the misconception depends on the formulation of the item.

Methodology

Measures and participants

Starting from a mathematics achievement test developed in 2011 by the Italian National Institute for the Evaluation of Educational System (INVALSI), we have developed three further Math achievement tests. These tests with the original one (F1, F2, F3, F4) were administered to a probabilistic sample (stratified by SES) consisting of 1647 students attending grade 8 (14-year old students, on average). Each test covers geometry, arithmetic, algebra and probability, in a range of difficulty from those that require simple mathematical operations to those that require complex thinking. Math ability has been estimated by Rasch model (Rasch, 1960) and scaled in an empirical range equal to [-4; +4], i.e. the latent trait along which items and persons are scaled depending on their difficulty and their ability, respectively. Each form consists of two groups of items, put in the same order in each form: 1) sixteen invariant (anchoring) items across forms used to equate them; and, 2) seventeen items with the same question intent but different item format or formulation.

Analytic strategy

All tests were administered by means of a spiralling process to students within the same classrooms. This spiralling process typically leads to comparable, randomly equivalent groups taking different forms because “the difference between group level performance on the two forms is taken as a direct indication of the difference in difficulty between the forms” (Kolen & Brennan, 2004, p.13). Nevertheless, the existence of a remarkable relationship between item formulation and its psychometric functionality is widely and well-known as well as the impact of this relationship on information quality about estimated students’ abilities/competencies in mathematics. All forms were equated to compare scores earned on different forms (consisting of different formulation of the same item, i.e. with the same math content and the same question intent) by different subjects. This comparison allows quantifying the effect of each variation in item formulation on its probability of being answered correctly (Bolondi, Cascella, Giberti, 2018). To compare items we explored their difficulty and goodness-of-fit via the weighted mean square error (MNSQ) as well as their characteristic curves (ICCs) that graphically link the probability of a correct answer to students’ ability. Through them, we observed and quantified how much different formulations affect the probability of a correct answer and how this relationship vary depending on students’ mathematics ability level.

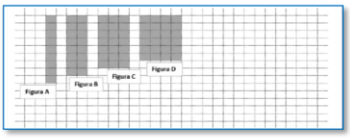
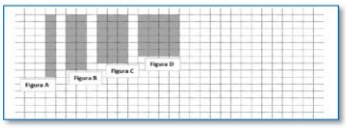
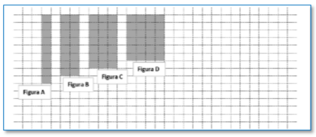
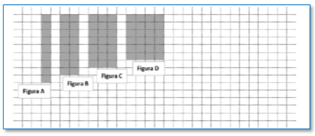
In this study, we assumed that the probability of a correct answer of course - and coherently with Rasch model theoretical assumptions - depends on students’ ability compared to item difficulty but also that variations in the probability of a correct answer are due to the activation of different solving strategies.

All analyses were performed by both ConQuest 4.0 and Rumm2030. Both of them provide measures from the Rasch framework, although with a different terminology (e.g., item difficulty is reported as Item Delta by ConQuest and as Locn by Rumm2030). ConQuest 4.0. adds also some measures from the Classical Test Theory.

The case study

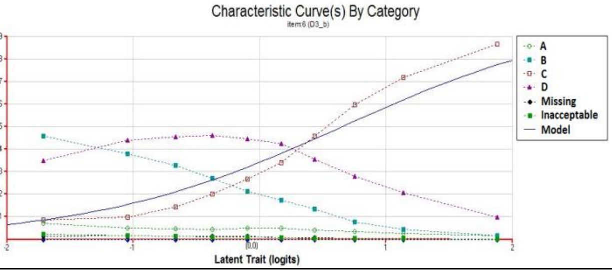
Table 1 shows four different formulations of the same item (D14) developed to explore students' misconception about the relationship between perimeter and area.

Table 1. Item formulation in Booklet F1, F2, F3, and F4

Booklet F1	**Booklet F2**
D14. Look at the following figure:  Complete the following statements by choosing, for each of them, the expression that makes the statement correct **a.** Areas of figures .... [do not change /increase at each step /decrease at each step] **b.** Perimeters of figures .... [do not change /increase at each step /decrease at each step]	**D14.** Look at the following figure:  Complete the following statements by choosing, for each of them, the expression that makes the statement correct **a.** Perimeters of figures .... [do not change /increase at each step /decrease at each step] **b.** Areas of figures .... [do not change /increase at each step /decrease at each step]
Booklet F3	**Booklet F4**
• Look at the following figure:  • Which of the following statement is true? A. Areas of the figures do not change B. Areas of the figures double at each step C. Perimeters of figures do not change D. Perimeters of figures increase at each step	• Look at the following figure:  • Which of the following statement is true? A. Perimeters of figures do not change B. Perimeters of figures increase at each step C. Areas of the figures do not change D. Areas of the figures double at each step

Booklet F3 contained (without any modifications) the original version of this item, administered by INVALSI, in 2012, to a sample of 39668 students. Data analysis of the INVALSI large-scale study shows a little misfit for this item, even though weighted MNSQ is acceptable (0.91). In particular, the graph below (Table 2) shows that the model overestimates the probability of choosing the correct answer of medium-low ability levels and underestimates it for higher ones. The percentage of correct answers is really low (36.6%) for this item that can be simply solved by counting squares and segments for area and perimeter, respectively. Item difficulty is confirmed by Rasch parameter and its high discrimination. 21% of students choose distractor B, probably because students noticed only that the base of the rectangles increases. The most attractive option (35.41%) is D (i.e., perimeters increase at each step).

Table 2. Item functionality observed in main study



Cases for this item 39668	Discrimination 0.52
Weighted MNSQ 0.91	Item Delta 0.66
Label	**% of tot**
A	4.39
B	21.65
C (correct)	36.60
D	35.41
Missing (answer not provide)	1.00
Inacceptable answer (excluded)	0.95

Based on those empirical findings, we hypothesized that students (especially with low and medium ability) compared figures and noticed that their areas increase and, subsequently, they fell into mistakes about perimeter because of misconception that perimeter increases when area increases, and vice versa - i.e., if area decreases, also perimeter decreases (Tirosh and Stavy, 1999; Tsamir & Mandel, 2000; D'Amore and Fañdino Pinilla, 2005; Machaba, 2016).

This misconception seems to be really strong because students simply had to count squares and segments to provide a correct answer. In this perspective, a recent research argued that the order of the options (item about perimeter followed by item about area, or vice versa) can influence students' answers (D'Amore and Fañdino Pinilla, 2005). In line with this argumentation, we modified INVALSI item by inverting item order (i.e., item about perimeter followed by item about area) and we also tried to estimate the effect of different item format, i.e. multiple choice versus open ended format (Table 1).

Results

This paragraph reports on item functionality explored by ConQuest 4.0 (Table 3) and via the graphical inspection of ICCs, plotted by RUMM2030 (Table 4 and Table 5). In table 5, we added also distractor plots. Similarly to the ICCs, distractor plot expresses the probability of choosing an answer option given by its relative difficulty, i.e. its difficulty compared to students' ability.

Table 3. Item functionality in F1, F2, F3 and F4.

F1				F2			
item:29 (D14_a) Discrimination 0.38 Weighted MNSQ 0.94		**item:30(D14_b)** Discrimination 0.57 Weighted MNSQ 0.96		**item:28(D14_a)** Discrimination 0.49 Weighted MNSQ 0.93		**item:29(D14_b)** Discrimination 0.41 Weighted MNSQ 0.93	
Label	% of tot	Label	% of tot	Label	% of tot	Label	% of tot
Wrong	9.09	Wrong	38.52	Wrong	47.55	Wrong	14.95
Correct	84.45	Correct	55.02	Correct	47.30	Correct	78.92
Missing	6.46	Missing	6.46	Missing	5.15	Missing	6.13
F3				**F4**			
item:29(D14) Discrimination 0.60 Weighted MNSQ 0.83				**item:29(D14)** Discrimination 0.55 Weighted MNSQ 0.92			
Label	% of tot			Label	% of tot		
A	3.41			A (correct)	45.64		
B	13.14			B	39.82		
C (correct)	42.34			C	2.91		
D	38.20			D	9.62		
Missing	2.92			Missing	2.01		

Table 4. ICCs of D14a and D14b in F1 (on the left) and F2 (on the right)

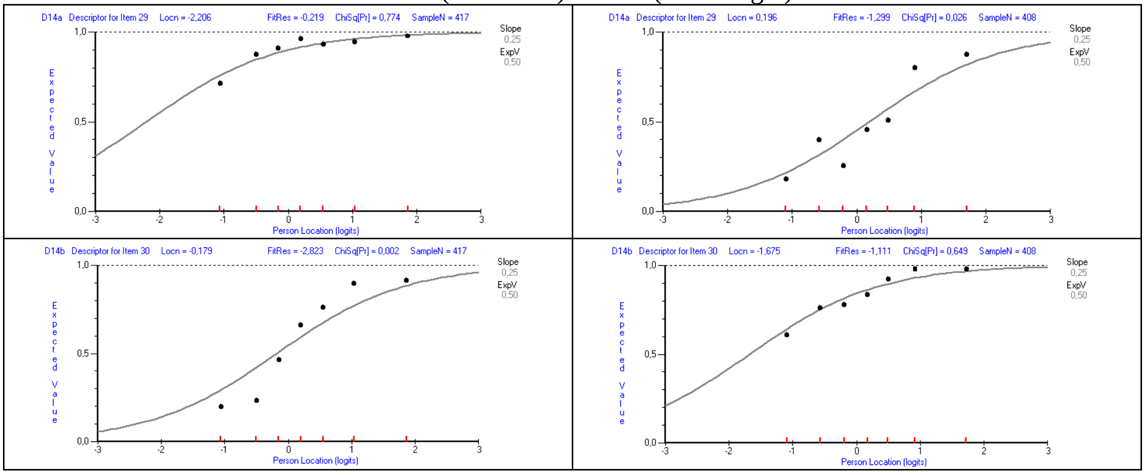
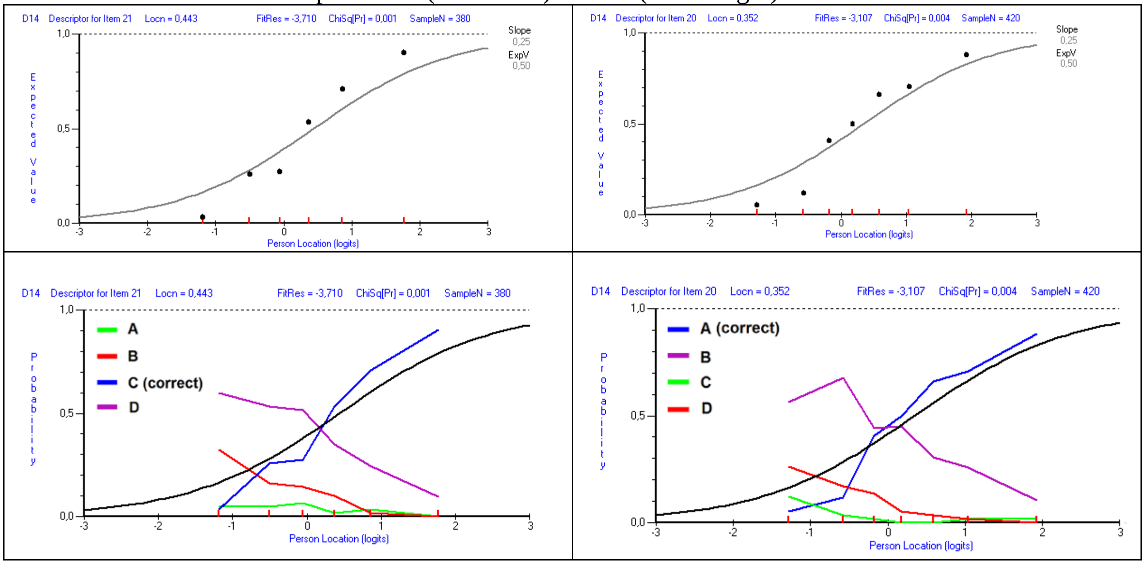


Table 5. ICC and distractor plot in F3 (on the left) and F4 (on the right)



When students are asked about perimeter and then about area (F2) in lieu of the reverse (F1), the percentage of correct answers decreases by around 10%. This difference decreases (and becomes equal to around 3%) by comparing multiple choice items (in F3 and F4), suggesting that the effect of item order can be explored better by splitting the question into two different items and by using an open-ended item format.

The graphical inspection of item characteristic curves adds some interesting information about the effect of our formulations on the probability of a correct answer depending on students' ability level. First at all, ICCs plotted for items D14a and D14b in F1 and F2 (Table 4) show that, irrespective of item order, items about areas show strong guessing effect and low discrimination. These features are strongly interrelated: guessing effect reflects the fact that the Rasch model predicts for students with lower ability level a high probability of a wrong answer but observes a high percentage of correct answers (i.e., higher than its expectations) and this is in line with the fact that this item does not discriminate enough between students with higher ability and lower ability. The strong guessing effect and the low discrimination in the two tasks focused on area means that this question is pretty easy for all the students, and also students with lower ability levels notice that areas are increasing. However, there is a little difference between the difficulty: also in this case, the area task is easier if presented first (F1: locn=-2.206) compared to the same item presented after perimeter (F2: locn=-1.675). Variation in item difficulty can be disclosed also between D14b in F1 (locn=-0.179) and D14a in F2 (locn=0.196), as further proof of the fact that item order has an effect on the probability of completing the same item successfully and thus suggesting that item order affects students' solving strategies, and this may be due to a different influence of misconceptions as shown in D'Amore and Fañdino Pinilla (2005). Finally, compared to item 14a in F2, the item 14b in F1 shows a significant over-discrimination.

Modification in item order and in item format produces much more information than just modification in item order. In fact, item functionality in F3 and F4 are very similar in their difficulty (slightly higher than zero in both cases) and in their goodness-of-fit (close to 1 and thus good) as well as in their distractors' functionalities. In particular, the distractor "Areas of the figures do not change" (option A and C in F3 and F4 respectively) was not attractive for all ability levels; the option "Areas of the figures double at each step" (B and D, in F3 and F4 respectively) was slightly more attractive; finally, few differences – that might be deepened by interviewing students – were disclosed by comparing options C and D with options A and B, respectively in F3 and F4.

Conclusions

Our experiment confirms the existence of the misconception about the relationship between area and perimeter and the anchoring technique shows that the dimension of the group of students influenced by the misconception depends on the formulation: it is more difficult to correctly answer the item if the formulation leads the student to focus first on the area and then on the perimeter. As a by-product, we see that, from a statistical point of view, the stronger the misconception acts, the weaker is the fit of the item with respect to the theoretical model.

References

Bolondi, G., Branchetti, L., & Giberti, C. (2018). A quantitative methodology for analyzing the impact of the formulation of a mathematical item on students learning assessment. *Studies in Educational Evaluation, 58*, 37-50.

Bolondi, G., Cascella, C., Giberti, C. (2018, in press). Formulazione della domanda e funzionalità psicometrica. In P. Falzetti (Ed.), *I dati INVALSI: uno strumento per la ricerca.* Milano: Franco Angeli.

D'Amore, B., & Fandiño Pinilla, M. I. (2005). Relazioni tra area e perimetro: convinzioni di insegnanti e studenti. *La matematica e la sua didattica, 2*, 165-190.

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer: New York

Machaba, F. M. (2016). The concepts of area and perimeter: Insights and misconceptions of Grade 10 learners. *Pythagoras, 37*(1), 1-11.

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

Tirosh, D., & Stavy, S. (1999). Intuitive rules: A way to explain and predict students reasoning. *Educational Studies in Mathematics, 38*, 51–61.

Tsamir, P., & Mandel, N. (2000). The intuitive rule same A - same B: The case of area and perimeter. In T. Nakahara, & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4) (pp. 225–232). Hiroshima: PME.