DOI: 10.1478/AAPP.99S1A4

AAPP | Atti della Accademia Peloritana dei Pericolanti
Classe di Scienze Fisiche, Matematiche e Naturali
ISSN 1825-1242

Vol. 99, No. S1, A4 (2021)

WHAT CAN WE LEARN FROM LARGE-SCALE SURVEYS ABOUT OUR STUDENTS LEARNING OF MATHS?

GIORGIO BOLONDI *

ABSTRACT. Large-scale surveys on Maths and Science learning (such as OCSE-Pisa, IE-TIMMS, TIMMS Advanced and the INVALSI in Italy) have a strong influence on public opinion in all countries and, in a top-down process, on decisions by policy-makers and administrative stakeholders, organization of the education system, official curricula, actual curricula, and the didactic choices of teachers. This process is activated principally by a mechanism of comparison and ranking which is implicit in the published results of these surveys – a mechanism which induces effects which are not always positive (such as teaching to test). This study sets out to show, with some examples from the ongoing research, how it is possible to analyse macro-phenomena revealed by the survey results with conceptual tools of mathematics education, in order to look beyond the statistical data of individual students' performances or of the sample group as a whole. The quantitative analytical tools used in processing the information collected in these surveys can be used to suggest valuable clues in understanding the nature and origins of common misconceptions and difficulties, and how these are linked with didactic practices. The first case that we consider regards the answering of questions which highlight a strong difference between male and female students (the so-called gender gap): which questions are these and why? The second case is the analysis of some INVALSI questions through which it is possible to quantify a well-known didactic phenomenon: the "loss of meaning" in algebra. The second case regards the answering of questions which highlight a strong difference between male and female students (the so-called gender gap): which questions are these and why? The third case that we present shows how it is possible to study how students' behaviour (and their ability to find problem-solving strategies, for example) is influenced by the layout and wording of the question. These and other examples show how mathematics education can greatly benefit from the use of mixed methods (quantitative/qualitative) in surveys and research.

1. The impact of large-scale surveys of the learning of Mathematics

Large-scale surveys of Mathematics and Science learning (such as OECD-Pisa, IEA-TIMMS, IEA-TIMMS Advanced, and the national assessments such as INVALSI in Italy) have a strong influence on public opinion in all countries. With a top-down process, they impact on the decisions taken by policymakers and administrative stakeholders, on the organization and the general architecture of the educational system, on official intended

curricula and on actually implemented curricula, and ultimately on the didactic choices of the teachers. This process is activated principally by a mechanism of comparison and ranking which is implicit in the published results of these surveys – a mechanism which induces effects which are not always positive (such as *teaching to test*). Standardised Testing (Morris 2011) in mathematics may play a crucial role in this diagnostic process, even if many theoretical issues must be discussed. An issue at the centre of hard epistemological and ideological debates is how to integrate theoretical frameworks, results, methods, and tools of standardized assessments – that are designed in order to impact at a systemic level – into the local actions of teachers and schools (Looney 2011). In fact, teachers do not need rankings, they need operative and interpretative tools. Following (De Lange 2007), a bottom-up process would be helpful for integrating this top-down impact of large-scale assessments results (which are by their nature summative) into with classroom practices, and in particular with formative assessment activities (Looney 2011). For a general discussion on the use of large-scale assessment data in research in Mathematics education one can refer to De Lange (2007) and Meinck *et al.* (2017).

This study sets out to show, with three examples from the ongoing research, how it is possible to analyse macro-phenomena revealed by the survey results, by means of the conceptual tools developed by mathematics education research, in order to look beyond the statistical data of individual students' performances or of the sample group as a whole. The quantitative analytical tools used in processing the information collected in these surveys can be used to suggest valuable clues in understanding the nature and origins of common misconceptions and difficulties, and how these are linked with didactic practices. The goal is to shift the focus, from *measuring to understanding*.

Our general mixed-method strategy is as follows: we start from a macrophenomenon emerging from the results of a large-scale assessment; we try to interpret it with a theoretical construct or an empirical finding of research in Mathematics education; we validate the interpretation through a qualitative methodology based on case studies and/or observation. This strategy may also help in suggesting, when possible and suitable, didactic interventions.

2. Situation 1: Gender gap in Mathematics achievements

The first example of a research interplaying quantitative data coming from large-scale surveys and qualitative inquiries is about the *gender gap*.

Gender differences in mathematics performances, which are reported by both international and national large-scale surveys, have been debated in several studies and many researches focused on the determinants of gender-gap (Forgasz *et al.* 2010). This issue has been deeply studied, in particular by researchers in Mathematics education, from both a qualitative and a quantitative point of view. (Winkelmann *et al.* 2008) identifies both external and internal factors as possible factors determining this gap. Internal factors include for instance biological variables, but international surveys have revealed that gender gap in maths differs across countries (OECD 2015; Contini *et al.* 2017). Hence internal factors must be accompanied with other explanations connected to social and cultural factors, related to the context in which the students live. In this perspective, many researchers highlight the fact that in more gender-equal cultures this gap tends to disappear (Guiso *et al.* 2008; OECD 2015; Cascella 2017; OECD 2017). Furthermore, beliefs of teachers and parents about

boys and girls math abilities and gender stereotypes play an important role in students'self-perception and then have a huge influence on their performances (Jacobs and Bleeker 2004; Riegle-Crumb 2005; Fryer and Levitt 2010). Other studies have also outlined evidence of differences in metacognitive aspects related to maths: girls tend to be more influenced by math anxiety and display less math self-efficacy (Pajares 2005; OECD 2015; Cargnelutti *et al.* 2016; OECD 2017). Italy is one of the countries with the deepest gender gap in Mathematics.

Our research has been focused at its beginning on the distribution of the gender gap along the ability scale, as measured by a standardized assessment. The following is a typical distribution of boys' and girls' abilities (INVALSI test, Grade 06, 2013 - the x-axis represents the ability):

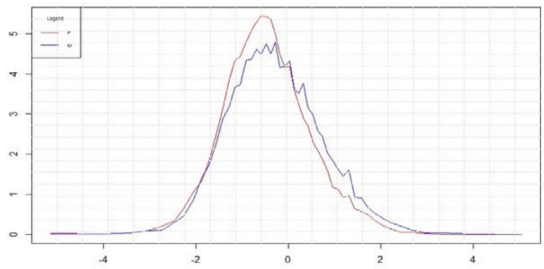

FIGURE 1

The gender gap is essentially a lack of top-performer girls. By the way, there is no gender gap in the distribution of school marks- the gap is detected only by *external* assessments.

The second step of the research has been the analysis of 1,400 items administered, from 2008 on, by the INVALSI. In fact, the difference between boys' and girls' performances is not the same for all items. We used different statistical tools (Bolondi, Cascella and Giberti, 2017; Cascella, Giberti and Bolondi, 2020) for pointing out those items for which the behaviour of boys and girls is particularly different. When an item response modelling is employed to analyse data, it is important to verify that observed data are consistent to the theoretical assumptions underlying the model. From the methodological point of view, the general issue that we tackled was how to use quantitative educational data to understand causes of item misfit. In fact, our research hypothesis is that a moderate misfit need not necessarily to be interpreted as a limitation (of the test or even of the choice of the model) but as a potential source of information both about the actual construct measured by the test, and the misfitting item itself.

As a result of this approach, it turned out that the gender gap is created by those items that seem to be *far from the classroom practice*. This is coherent with the previously observed fact that there are less top-performer girls. From a more strict didactic perspective, girls seem to be more sensible to what is called the *didactic contract* (Brousseau 1988, 2017; Bolondi *et al.* 2018b).

In fact, a qualitative research with a large sample of teachers, issued from this analysis, allowed to make explicit some beliefs of Italian teachers with respect to *which are the*

features characterizing a "good" student in mathematics: the adjectives used in order to describe boys are different from those used in order to describe girls. Features like *diligente, attenta, studiosa* are commonly used for girls but they are not used for boys. It seems that there are stereotypes about gender roles and features which are present in the classrooms, implicit in teachers' habits, which are coherent with the belief that girls may be good, but just because they are assiduous, careful, and that they cannot be really outstanding. There is a well known phenomenon called *Glass Ceiling effect* which is the macro-social effect and counterpart of this stereotype.

3. Situation 2: Loss of meaning of algebraic symbols

The second example is related to a general didactic problem: *the loss of meaning* of algebraic symbols.

The learning of algebra is a central issue in mathematics education. It is situated into a new cognitive scenario where procedural thinking and relational thinking are closely interfaced. It involves since its beginning *modelization* (for instance, of real world situations) and *generalization* (for instance, of arithmetic relations): algebra may be seen as a turning point in the student's educational path that entails new forms of thinking and conceptual organization; it marks the move from pure arithmetical thinking to advanced mathematics. Algebraic thinking is characterized by the introduction of a formal symbolic language whose relationship with the meaning is crucial (Arcavi 1994; Capraro and Joffrion 2006; Kieran 2016). This introduction is spoiled by an intrinsic difficulty: the basic objects of the algebraic discourse are needed in a given step of the curriculum, hence must be presented to students of a given age, but they can be properly defined only in a formal setting which is non accessible by students of that age (Bolondi *et al.* 2020). The introduction of formal language therefore involves important cognitive difficulties on the part of the student; thus, a common outcome of school algebra is a mechanical use of signs that are void of meaning. The main educational issue is to prevent such a manipulative use of letters and insert algebra in the mathematical curriculum as a true instrument for new forms of thinking and generalizations. Arzarello *et al.* (2006) describe the obstacles and the difficulties in the learning of algebra, outlining how the use of algebraic formalism causes a dissonance between procedural thinking conveyed by natural language and relational thinking conveyed by the symbolic one. There is a loss of semantic control of algebraic operations that result in a meaningless transformation of signs.

Our research was aimed at *quantifying* this loss of meaning in the use of algebraic symbols. We individuated two items, form the INVALSI large-scale assessment in Grade 08 and Grade 10, which clearly describe what happens. We analysed data consisting of the answers given by samples of students of Grades 08 and 10 from across the country, involving 28,361 students in 1,312 classes representative of 586,790 students, and 50,838 students in 2,302 classes representative of 527,318 students, respectively.

The analysis shows that the semantic control of literal expressions, that are here used as a tool of generalization and modelling, decreases between the two educational levels, despite two years of school that are devoted to the development of algebraic language: the difficulty increases and the value of the Item delta inside the tests jumps from -0.42 to 0.53. The number of correct answers decreases from 58.3% to 38.1%. Both items have good

D17. **La formula $L = L_0 + K \times P$ esprime la lunghezza L di una molla al variare del peso P applicato. L_0 rappresenta la lunghezza in centimetri "a riposo" della molla; K indica di quanto si allunga in centimetri la molla quando le si applica una unità di peso.**
Quale delle formule elencate si adatta meglio alla seguente descrizione:
"È una molla molto corta e molto dura (cioè molto resistente alla trazione)"?

- ☐ A. $L = 10 + 0{,}5 \times P$
- ☐ B. $L = 10 + 7 \times P$
- ☐ C. $L = 80 + 0{,}5 \times P$
- ☐ D. $L = 80 + 7 \times P$

FIGURE 2. Item D17, INVALSI test for Grade 08, 2011

D24. **La formula $l = l_0 + k \cdot P$ esprime la lunghezza l di una molla al variare del peso P applicato. l_0 rappresenta la lunghezza in centimetri "a riposo" della molla; k indica di quanto si allunga in centimetri la molla quando si applica una unità di peso.**
Quale delle formule elencate si adatta meglio alla seguente descrizione:
"È una molla molto lunga e molto resistente alla trazione"?

- ☐ A. $l = 15 + 0{,}5 \cdot P$
- ☐ B. $l = 75 + 7 \cdot P$
- ☐ C. $l = 70 + 0{,}01 \cdot P$
- ☐ D. $l = 60 + 6 \cdot P$

FIGURE 3. Item D24, INVALSI test for Grade 10, 2011

psychometric features: there is a good fitting with the model and both items discriminate between students. The percentage of missing answers (which usually are very few in a multiple choice test with no penalties for wrong answers, as in this case) increases from 4% to 11%, and this too can be seen as an evidence of a difficulty in giving a meaning to the task. An interesting evidence of the distribution of correct answers along the ability scale is that these "lost students" are high-ability students. For instance, in the eighth decile of the ability scale, the percentage of correct answers slumps from 80% to 50% when passing from Grade 08 to Grade 10; the percentage of correct answers in the ninth decile decreases from 85% to approx. 60%. Other interesting evidences can be drawn from the distribution of "wrong" choices.

4. Situation 3: Formulation of a task and performances of the students

The third example is related to a classical problem: how to compare different formulations of the same task? Which are the features (linguistic, graphical, organizational, figural...) which impact on students' understanding, strategies, actions and therefore performances? In fact, students are influenced by many facets of the formulation, when they face a task (a literature review can be found in Daroczy *et al.* (2015). The main issue is that this influence is difficult to study and to measure, since individual features of the students and context variables are clearly very important too. The optimal situation would be to administer two versions of the same task to the same group of students: but in fact, the answers to the

second version would be obviously influenced by the answers given to the first one. Several approaches have been used, in the last decades, to overcame this crucial issue (De Corte *et al.* 1985; Lepik 1990; Thevenot *et al.* 2007; Vicente *et al.* 2008).

Our research was aimed at developing a tool for measuring this impact, based on the use of large-scale assessment results, as a tool for researchers. The goal was to design a quantitative methodology integrating the existing research approaches, in order to address the point of measuring and differentiating the impact on the students' performances due to a variation in the formulation of a task. The tool is based, from the statistical point of view, on an anchoring technique and it has been validated with an articulated strategy (Bolondi *et al.* 2018a)

It works as follows- we sketch the strategy without entering technical details. We start from the results of a large-scale assessment for the grade and the context we are interested in- for instance, Italian students of Grade 08. We select in the test a subset of items that are representative of the construct we are dealing with-in our case, mathematics literacy as defined by the National curricula. These items are representative both qualitatively (they cover key aspects of the curriculum) and quantitatively (they give a measure of the ability of the students which is coherent with the measure of the whole test). This subset of items will be the *core test* CT of the measurement. We prepare two (or even more) variants of the formulation of an item, V_1 and V_2. Then we prepare two tests T_1 and T_2, the first composed by CT and V_1, the second composed by CT and V_2. We select a sample of students, representative of the population we are interested in, and we divide randomly the population in two parts P_1 and P_2. We administer T_1 to P_1 and T_2 to P_2. Then, we measure with a Rasch-model procedure the ability of each student (of both P_1 and P_2) with CT, and then the difficulty of V_1 and V_2 considering the abilities of the students of P_1 and P_2 respectively. We define a set of coherence criteria to be verified, which guarantees the applicability of the method. The outputs of the technique are:

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

With this technique, we were able to measure the impact of several typologies of variations in the formulation: size of the numbers involved, order of factor in an arithmetical operation, position of a picture in the text, order of words, linguistic complexity (Bolondi *et al.* 2018a).

5. Conclusions

Large-scale assessment can be a helpful tool for understanding dynamics and outputs of teaching-learning processes. They need to be interpreted with theoretical lenses and integrated with the results of the research in Mathematics education. Research paradigms in Mathematics education are mainly qualitative (Hart *et al.* 2009) and for many didactic phenomena there are detailed case-study descriptions and theoretical frameworks. Nevertheless, quantification of phenomena can be useful and sometimes necessary for on-the-field teachers, and mixed-method research methodologies can be fruitful for researchers. In this paper we presented three examples of how frameworks, released items and quantitative

results from large-scale assessment can be interfaced with qualitative results and research methodologies.

References

Arcavi, A. (1994). "For the Learning of Mathematics". *For the Learning of Mathematics* **14**(3), 24–35.

Arzarello, F., Bazzini, L., and Chiappini, G. (2006). "A Model for Analysing Algebraic Processes of Thinking". In: pp. 61–81. DOI: 10.1007/0-306-47223-6_4.

Bolondi, G., Branchetti, L., and Giberti, C. (2018a). "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. DOI: 10.1016/j.stueduc.2018.05.002.

Bolondi, G., Ferretti, F., and Giberti, C. (2018b). "Didactic Contract as a Key to Interpreting Gender Differences in Maths". *ECPS - Educational Cultural and Psychological Studies* **18**, 415–435. DOI: 10.7358/ecps-2018-018-bolo.

Bolondi, G., Ferretti, F., and Maffia, A. (2020). "Monomials and polynomials: the long march towards a definition". *Teaching Mathematics and its Applications* **39**(03), 1–12. DOI: 10.1093/teamat/hry015.

Brousseau, G. (1988). "Le contrat didactique: le milieu". *Recherches en Didactique desMathématiques* **9**(03), 309–336. DOI: 10.1093/teamat/hry015.

Brousseau, G. (2017). "The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics". In: *Theory of mathematics education ICME 5 – topic area and miniconference Adelaide, Australia*. Ed. by H. Steiner. Bielefeld: Institut für Didaktik der Mathematik der Universität Bielefeld, pp. 110–119.

Capraro, M. and Joffrion, H. (2006). "Algebraic Equations: Can Middle-School Students Meaningfully Translate from Words to Mathematical Symbols?" *Reading Psychology* **27**(2), 147–164. DOI: 10.1080/02702710600642467.

Cargnelutti, E., Tomasetto, C., and Passolunghi, M. C. (2016). "How is anxiety related to math performance in young students? A longitudinal study of Grade 2 to Grade 3 children". *Cognition & emotion* **31**(4), 1–10. DOI: 10.1080/02699931.2016.1147421.

Cascella, C. (2017). "Exploring the relationship between social roles in daily life and achievement gap between boys and girls in maths: Empirical evidences from Italian primary school". In: *11th annual International Technology, Education and Development Conference Proceedings*. Valencia: IATED, pp. 9832–9841. DOI: 10.21125/inted.2017.2339.

Contini, D., Tommaso, M., and Mendolia, S. (2017). "The gender gap in mathematics achievement: Evidence from Italian data". *Economics of Education Review* **58**, 32–42. DOI: 10.1016/j.econedurev.2017.03.001.

Daroczy, G., Wolska, M., Meurers, D., and Nuerk, H.-C. (2015). "Word problems: A review of linguistic and numerical factors contributing to their difficulty". *Frontiers in psychology* **66**(6), 314, [14 pages], 348. DOI: 10.3389/fpsyg.2015.00348.

De Corte, E., Verschaffel, L., and Win, L. (1985). "Influence of Rewording Verbal Problems on Children's Problem Representations and Solutions". *Journal of Educational Psychology* **77**(4), 460–470. DOI: 10.1037/0022-0663.77.4.460.

De Lange, J. (2007). "Large-scale assessment and mathematics education". In: *Second handbook of research on mathematics teaching and learning*. Ed. by F. J. Lester. Charlotte, NC: Information Age Publishing, pp. 1111–1142.

Forgasz, H., Becker, J., Lee, K., and Steinthorsdottir, O. (2010). *International perspectives on gender and mathematics education*. Charlotte, NC: Information Age Publishing.

Fryer, R. G. and Levitt, S. D. (2010). "An empirical analysis of the gender gap in mathematics". *American Economic Journal: Applied Economics* **2**(2), 210–240.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). "Culture, Gender, and Math". *Science* **320**.

Hart, L., Smith, S., Swars Auslander, S., and Smith, M. (2009). "An Examination of Research Methods in Mathematics Education (1995-2005)". *Journal of Mixed Methods Research* **3**, 26–41. DOI: 10.1177/1558689808325771.

Jacobs, J. and Bleeker, M. (2004). "Girls' and boys' developing interests in math and science: Do parents matter?" *New directions for child and adolescent development* **2004**, 5–21. DOI: 10.1002/cd.113.

Kieran, C. (2016). "Research on the learning and teaching of algebra". In: *The Second Handbook of Research on the Psychology of Mathematics Education*. Ed. by A. Gutiérrez and P. Boero. Rotterdam: Sense Publishers, pp. 11–50. DOI: 10.1007/978-94-6300-561-6_3.

Lepik, M. (1990). "Culture, Gender, and Math". *Educational Studies in Mathematics* **21**, 83–90. DOI: 10.1007/BF00311017.

Looney, J. W. (2011). "Integrating Formative and Summative Assessment: Progress Toward a Seamless System?" *OECD Education Working Papers* (58). DOI: 10.1787/5kghx3kbl734-en.

Meinck, S., Neuschmidt, O., and Taneva, M. (2017). "Workshop Theme: "Use of Educational Large-Scale Assessment Data for Research on Mathematics Didactics"". In: *Proceedings of the 13th International Congress on Mathematical Education*. Ed. by G. Kaiser. Cham: Springer International Publishing, pp. 741–742. DOI: 10.1007/978-3-319-62597-3_132.

Morris, A. (2011). "Student Standardised Testing: Current Practices in OECD Countries and a Literature Review". *OECD Education Working Papers* (65). DOI: 10.1787/19939019.

OECD (2015). "The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence", 180. DOI: 10.1787/9789264229945-en.

OECD (2017). *PISA 2015 Assessment and Analytical Framework*, p. 260. DOI: 10.1787/9789264281820-en.

Pajares, F. (2005). "Gender Differences in Mathematics Self-Efficacy Beliefs". In: *Gender differences in mathematics: An integrative psychological approach*. Ed. by G. A. M. and J. Kaufma. Cambridge: Cambridge University Press, pp. 294–315.

Riegle-Crumb, C. (2005). "The cross-national context of the gender gap inmath and science". In: *The social organization of schooling*. Ed. by L. Hedges and B. Schneider. New York: NY: Russell Sage Foundation, pp. 227–243.

Thevenot, C., Devidal, M., Barrouillet, P., and Michel, F. (2007). "Why does placing the question before an arithmetic word problem improve performance? A situation model account". *Quarterly journal of experimental psychology (2006)* **60**(01), 43–56. DOI: 10.1080/17470210600587927.

Vicente, S., Orrantia, J., and Verschaffel, L. (2008). "Influence of situational and conceptual rewording on word problem solving". *The British journal of educational psychology* **77**(04), 829–848. DOI: 10.1348/000709907X178200.

Winkelmann, H., Heuvel-Panhuizen, M. van den, and Robitzsch, A. (2008). "Gender differences in the mathematics achievements of German primary school students: Results from a German large-scale study". *ZDM* **40**(04), 601–616. DOI: 10.1007/s11858-008-0124-x.

* Libera Università di Bolzano
Scienze della Formazione
Viale Ratisbona 16, 39042 Bressanone, Bolzano, Italy

Email: giorgio.bolondi@unibz.it

Paper contributed to the international workshop entitled “New Horizons in Teaching Science”, which was held in Messina, Italy (18–19 november 2018), under the patronage of the *Accademia Peloritana dei Pericolanti*
Manuscript received 17 June 2020; published online 30 September 2021