Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 112 (2014) 180 – 189

Procedia
Social and Behavioral Sciences

International Conference on Education & Educational Psychology 2013 (ICEEPSY 2013)

The evaluation of mathematical competency: elaboration of a standardized test in Ticino (Southern Switzerland)

Alberto Crescentini*, Giovanna Zanolla

Department of Training and Learning, University of Applied Sciences and Arts of Southern Switzerland
Piazza San Francesco, 19 – 6600 Locarno, Switzerland

Abstract

Since 2010 a project with the aim of producing and administering a standardized test (Woolfolk, 2007) to evaluate mathematical competencies in the fourth class of primary school has been started in Ticino. In order to produce the test several steps was necessary: a team has first identified the areas of the mathematical program to be tested, second a group composed by primary and lower secondary school teachers, discipline experts and teachers of mathematical didactics has produced items coherent with the aim and with the characteristics of students and school programs, third the item produced were tested on a sample of students to evaluate the discriminative capacities of the items, fourth a preliminary analysis of the items was carried out, fifth the test was produced and administered to the whole population of students. In the fourth phase we used the classical Rasch model (1960) to evaluate and select the items and the software ConQuest supported in the analysis.
Every teacher received a report on his or her own class in which it is possible to identify the strengths and weaknesses of the class on each part of the tested program.
This is the first experience in Ticino in producing this kind of tests in the primary school.

"Keywords: Test competencies, item response theory, standardized tests, evaluation"

1. Introduction

1.1. Background

The term “mental test” is traditionally linked to the work of Cattell (1890) who used the experimental method, in order to measure cognitive processes. Since these pioneering works many researchers in the psychological field have been trying to build material instruments to evaluate invisible or latent dimensions (like intelligence

* Corresponding author. Tel.: +41-058-666-6835; fax: +41-058-666-6819. *E-mail address:* alberto.crescentini@supsi.ch.

doi:10.1016/j.sbspro.2014.01.1154

but also competencies). These efforts have produced some instruments that are aimed to give some information that can be helpful in understanding similarities and differences among people. In order to make such a kind of evaluation it is indispensable that each instrument is administered and corrected always in the same way.

Standardized tests are administered, scored and interpreted in a standardized manner, i.e. with the same directions, time limits and scoring for all (Woolfolk, 2007). In the psychometrical jargon the use of term standardized when referring to tests is the tautological cause of the test need to be standardized to give useful information and to be considered a test at all (De Battisti, Salini & Crescentini, 2006). Competency test used in school are finalized at assisting in the evaluation of student's attainment in a content of a certain subject area in a certain country and in a specific class (Boncori, 1993). The tests can be administered to students of different grade level, concern different subjects and be composed of multiple-choice questions, true and false questions and short answers or essay questions subsequently recoded according to specific rules. They can be standard-referenced or criterion-referenced. In the first case the scores are determined by comparing how well individuals achieved on the test to other individuals who took the same test. In the second case scores are compared to certain predetermined criterion (Popham, 2011).

Standardized testing is a highly controversial and well debated topic. If on the one side they are a relatively objective tool for measuring student achievement that consumes little class time and produces useful information on which both teachers, school administrators and policy makers can rely in order to assess and improve their classes or schools, on the other side according to some authors they only reveal students' knowledge during the very short timeframe in which the tests are administered (Boaler, 2003). Moreover some students may not do well in standardized tests independently from their abilities for reasons connected to anxiety or to the pressures of the time limit attached to the tests (Buck, Ritter, Jenson & Rose, 2010). According to others (Moses and Nanna, 2007) they reflect the inequities that already exist within schools rather than meaningful differences in intelligence, student learning and teacher effectiveness and advantage the students from higher socioeconomic statuses. Two other risks mentioned about the abuse of standardized tests are that students undergo hours of intense test preparation in the classroom (Barrier-Ferreira, 2008) ending up being obsessed and that teachers devote most of the classroom time and of resources to preparing students for the standardized test (the so-called phenomenon of "teaching to test") (Popham, 2011), this latter phenomenon has been frequently mentioned as one of the consequences of the international tests such as PISA.

Despite all the criticism, we believe that the introduction for the first time in the Ticino's primary school of a standardized test in mathematics can have beneficial effects. In our opinion a certain score in the test must not be read as a definitive absolution or conviction but on the contrary it can help teachers to individuate their pupils' weaknesses and the strengths and to concentrate their efforts on the former. In no case it can replace the traditional evaluation method used by teachers in their own class that is based on a mix of different kinds of evaluation and must take in account specificities, history and characteristics of each student.

The primary school in Ticino is organized on a geographical basis. There are 9 districts (called "circondari") and in each district there is an inspector that is responsible for the quality of teaching. In the bigger schools there is a school manager that coordinates all the activities of the school. The nine inspectors are coordinated by a director of the infant and primary school.

In the primary schools of Ticino there is no tradition in using standardized tests. The learning process is followed with the aim that each student can reach his or her potentials and having a precise idea of the level of competency of the students has never considered relevant by the system in the last 20 years.

Recently some pressure in the direction of a more precise evaluation came from the federal organization due to a process of harmonization of compulsory education in Switzerland.

1.2. Purpose

During 2010 the responsible of the primary school of Ticino decided to evaluate the competencies in mathematics of the students at the end of the primary school. This evaluation needs to accomplish two main aims: on one hand it has to provide information about the system's performance and eventually about its gaps; on the other hand a feedback to each teacher about his or her class needs to be developed in order to give him or her the opportunity of implementing changes when they are necessary.

So there are three main objectives: developing a test on some previously defined math competencies; producing a feedback on the school system on math competencies; providing each teacher with a feedback report that might be of help in evaluating the performance of his or her students.

This paper is mainly devoted to describe the process of elaboration of the standardized test and to show the feedback report that was given to each teacher.

2. Methodology

2.1. Domains and test administration

As above specified our main goal was to develop a test able to assess the competency of the pupils of primary school in Canton Ticino in mathematics. Much of the efforts were spent in developing a set of more than 300 items concerning 6 competencies domains of mathematics and for this purpose a team of primary and lower secondary school teachers, discipline experts and teachers of mathematical didactics was assembled. The 6 dimensions (table 1) are only part of the school curriculum in Mathematics as the financial resources did not allow testing the whole program. This means that the choice has entailed the exclusion of other dimensions from the test. The main criterion that drive the choice is that the different domains need to be relevant in the specific scholastic year chosen.

Table 1. Chosen dimensions and areas

Dimensions	Areas
AR – Data and relations	SRD–Knowing, recognizing and describing
GM – Dimensions and measures	EA –Executing and use
GEO – Geometry	EA –Executing and use
GEO – Geometry	SRD –Knowing, recognizing and describing
NA– Numbers and calculating	AG –Arguing and Justifying
NC – Numbers and calculating	EA –Executing and use

In order to analyse the quality of the items on the basis of the items response theory (Lord, 1980) they were distributed over 10 booklets. Every pupils of a sample corresponding to half of the students of the 4th grade of primary school in Ticino (i.e. 1,683 pupils) was given two booklets with 60 items each at a distance of one week one another. The criterion followed in order to obtain reliable results for the quality of the items was that an item had to be administered to at least 200 to 300 pupils (table 2).

Table 2. Number of pupils per booklet (N = 1683)

Booklet	Number of pupils			Booklet	Number of pupils		
	T1	T2	Total		T1	T2	Total
A	163	157	320	F	164	159	323
B	163	165	328	G	161	163	324
C	163	163	326	H	157	153	310
D	153	162	315	I	157	160	317
E	159	164	318	K	162	154	316

2.2. Evaluation Model fit

Using the software ConQuest we analyzed the fit of the data to different items response models (one-dimensional and multidimensional Rasch model) then we compared the one-dimensional model "Mathematics" with a multidimensional model as well as with a one dimensional model with domains. Although the multidimensional model fitted the data significantly better than the one-dimensional (table 3), since the correlation between the six dimensions was quite high (they ranged from a minimum of 0.67 to a maximum of 0.84) we opted for fitting the data to a one-dimensional model and defining the domains at a second step. This choice is based on the assumption that one latent construct "Mathematics" exists and can be divided into different domains that are highly correlated and practically means that all item parameters are estimated basing on the one-dimensional model and the mean of all item parameters is constrained to 0. As table 4 shows the model with domains fits the data significantly better than the pure one-dimensional model.

Table 3. Comparison of the one-dimensional model with the multidimensional model

Models	Deviance	Number of parameters/df	Significance
One-dimensional model	102072.60	301	
Multidimensional model	101429.94	321	
Difference/comparison	642.66	20	$P < 0.001$

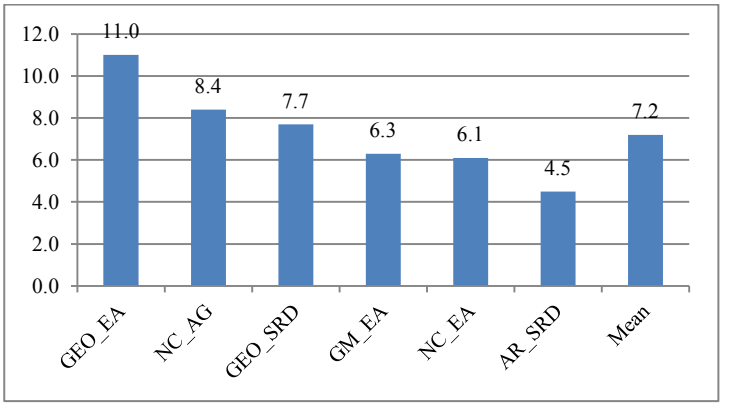
Table 4. Comparison of the one-dimensional model with the one-dimensional model with domains

Models	Deviance	Number of parameters/df	Significance
One-dimensional model	102072.60	301	
One-dimensional model with domains	101514.43	27	
Difference/comparison	558.17	274	$P < 0.001$

After having chosen the model, we evaluated the item quality looking at those ones which fitted the model better. Moreover we looked at the number of missing values per item and we deleted those with a high amount of

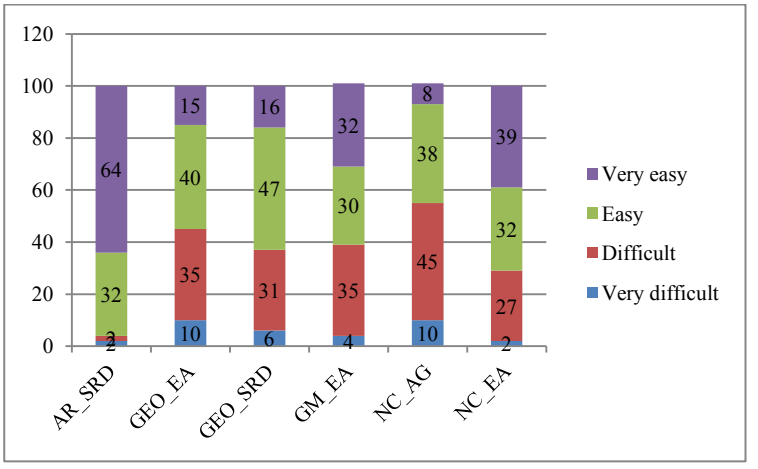
missing values as either they were extremely difficult or unclear. Figure 1 shows the average amount of missing values for dimension.

Figure 1. Average percentage of missing values per item and per domains



Since it can be expected that most pupils have average abilities, a test should consist of many items of average difficulty and a reduced amount of very difficult (solved correctly by 0 to 25% of the pupils) and very easy (solved correctly by 75 to 100% of the pupils) items. While the items in the domains GEO_EA, GEO_SRD and NC_AG are well distributed and cover a wide range of item difficulty, GM_EA and NC_EA should include some more difficult items and in AR_SRD the lack of difficult and very difficult items makes it very hard to identify pupils with high ability on this dimension (Figure 2).

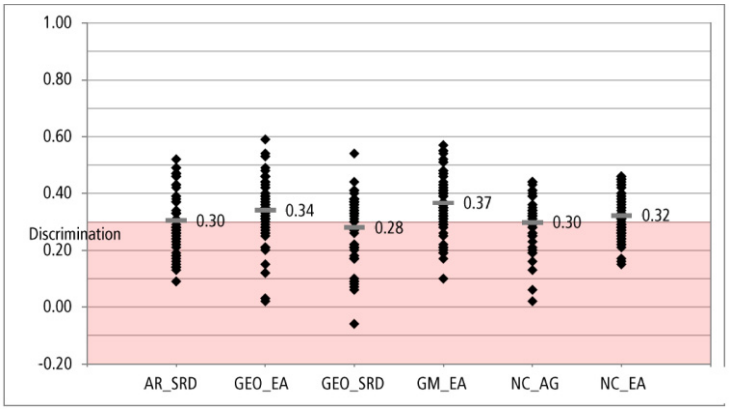
Figure 2. Distribution of item difficulty per dimension



We also examined the discrimination power of each item, which is given by the correlation of the item with the test row score. Items with a discrimination coefficient between 0.3 and 1 differentiate well between pupils that are strong and pupils that are weak in a certain skill. A coefficient near 0 indicates that the item does not

differentiate between strong and weak pupils. Items with a negative discrimination coefficient, which means that strong pupils solve the item less often correctly than weak pupils, were excluded (as figure 3 shows there is one in the dimension GEO_SRD). Possible reasons for a low discrimination include bad answer alternatives in multiple choice items, presence of a mistake in the answer key, very low o very high item difficulty.

Figure 3. Discrimination of items per domains



We further examined the infit (Weighted Mean Square MNSQ) that indicates the fit of an item to the Rasch model by analyzing the number of unexpected answers that differ from the prediction of the Rasch model. A value of 1 indicated a very good fit, a value smaller than 1 indicates that the item discriminates stronger than expected by the model and a value greater than 1 indicates that the item discriminates less than predicted by the model. All items had infit values between 0.7 and 1.3, which means acceptable but 47 had an infit values significantly different from the expected values and had to be excluded.

In addition to the infit value the item characteristic curve (ICC) of each item gives further information about the fix of the item to the Rasch model. Figure 4 shows an ICC (represented by a dotted curve) of an item that fits very well to the Rasch model (represented by the continuous line). On the contrary figure 5 shows the ICC of an item that does not fit very well to the Rasch model although its infit value is acceptable (MNSQ = 1.02). Such item had to be excluded.

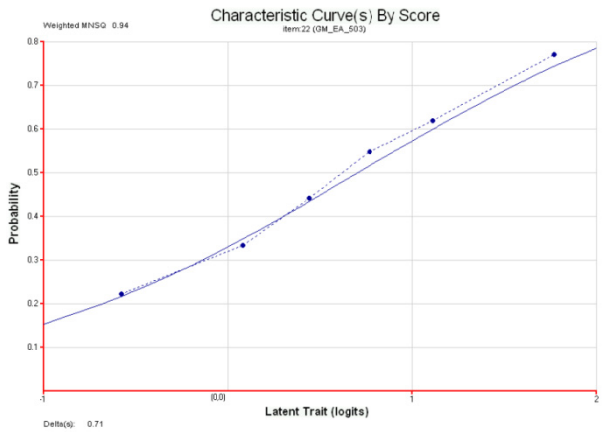


Figure 4. ICC of an item with good fit

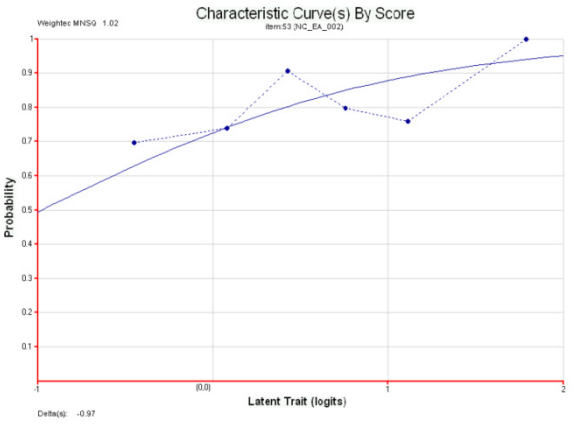
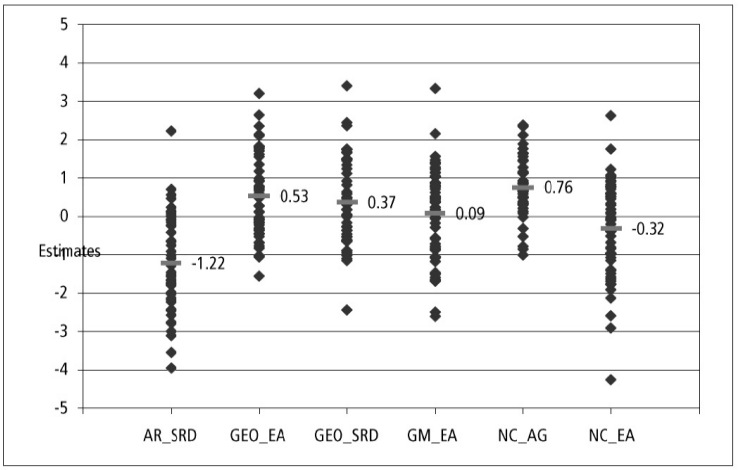


Figure 5. ICC of an item with bad fit

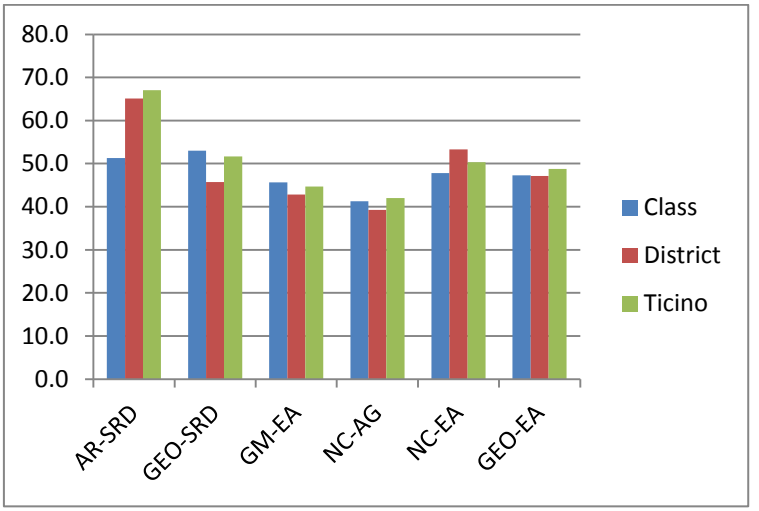
We finally examined the distribution of the parameter estimates and the differential item functioning (DIF).
A low parameter estimates indicates that an item is very easy to solve, while the contrary is true for a high parameter estimates. A parameter estimates of 0 refers to an item of average difficulty. It is important to have a broad range of items of different difficulties for every domains. Therefore we conclude that difficult items had been retained for the dimension AR_SRD and easy items had been retained for the domains GEO_EA, GEO_SRD and NC_AG (Figure 6).

Figure 6. Distribution of the item parameter estimates per domains



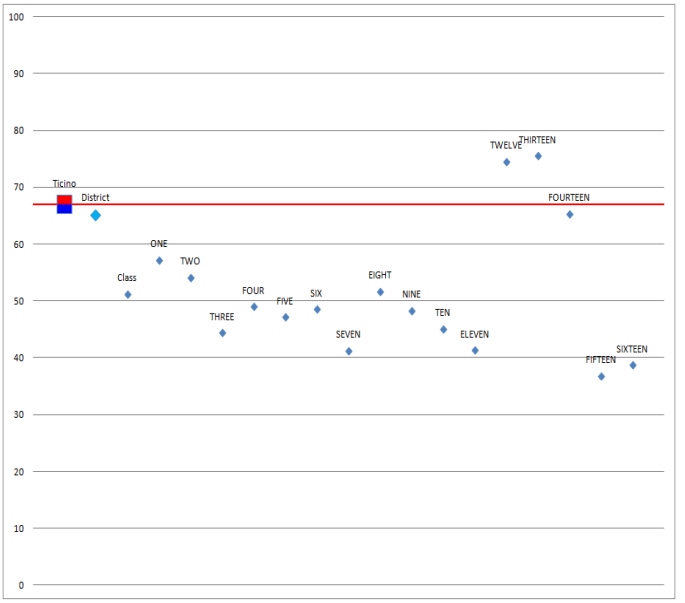
The best fitting 120 items (20 for each domains) were chosen to evaluate the whole population of students on the defined dimension. Some problems concerning the lack of difficult items for the domain AR_SRD and a lack of easy items for the domains GEO_EA and NC_AG remain and limit the possibility to describe the abilities of very able and very weak pupils on these dimensions. Additional items had to be developed to cover a broader range of difficulty or ability respectively but overall the result is acceptable.
Two different booklets were prepared, each one containing three domains to be tested. In the booklets the items were ordered from the less to the most difficult one. The order was decided using the previous analysis done on the sample of 1,683 pupils.
The whole population of students attending the 4th class of the primary school has been tested. It is composed of 2,947 students, 2,213 of whom Swiss and 734 have a different nationality. 1,935 speak Italian as mother tongue and 1,012 other languages. 196 classes of 136 schools have been tested. Moreover 385 students (about 13 % of the whole population) attend a class in which there are students of different ages, a typical condition in small communities.
Each teacher received a report on his or her own class with the explanation of the process followed and the description of the results. This report gives every teacher the opportunity to compare his or her class average performance score in each dimension with the average performance score of the whole population and of his or her district of belonging (figure 7 shows the scores related to an imaginary class X). The scale corresponding to each domain is given by the sum of the related items weighted by the difficulty coefficient resulted from the analysis of the first administration. The weighted- sum scores of the six dimensions were then normalized in order to assume values from 0 to 100 and facilitate reading data and graphics.

Figure 7. Example of graph in the class X's report



Furthermore for each domain considered in the report there is a graphic in which it is possible to identify the score of each student compared with the class, district and whole population average score (see the example related to an imaginary class in figure 8).

Figure 8. Example of a graph for AR-SRD domain in the class X's report



3. Results

Although the main aim of this contribution does not concern the specific results in the test we would like to conclude this work with a cue. The test puts into evidence that males obtain a higher (although not significative)

average score than females in the performance in every dimension. In literature there is a debate about differences between male and female in math performance related to the age and the result of absence of significative differences is coherent with other studies (Consorzio PISA.ch, 2011). A significative difference (P<.05) has been identified when for the general domain (normalized sum of the scores obtained in the six dimensions, is a general indicator of performance) subjects have been divided for quartiles of performance and male and female were confronted. In the fourth quartile there is a over representation of males, that is coherent with the work of Penner (2003).

Table 5. Quartiles comparison Male – Female in the general domain

	First quartile	Second quartile	Third quartile	Fourth quartile
Male	377	348	374	408
Female	352	381	354	320

In all the considered domains Swiss students obtain significative better results than foreigners (Table 6)

Table 6. Comparison Foreigner – Swiss students in the six domains

	AR-SRD	GEO-SRD	NC-EA	GEO-EA	GM-EA	NC-AG
Foreigners	62.6	48.2	47.5	44.6	40.8	38.6
Swiss	68.7	52.4	51.1	49.9	45.9	41.9

There are significative differences among the 9 districts in all the six domains considered. These differences seem to be connected more to the attitudes of the inspectors and to the history of the districts than to socio-economical variables.

In Ticino this is the first time that primary school teachers receive an “objective” evaluation of students’ performance. Most of the teachers sent us a positive feedback and at the moment a new project aimed at developing didactical instruments to answer to teachers’ questions and a project related to a different set of Mathematics dimensions are close to start.

4. Conclusions

The described research has involved the entire system of primary school in Ticino for more than three years. The objective evaluation of performance is an activity different from the ordinary work of teaching in primary school and it can produce mistrust by the teachers who may feel judged. The involvement of all the stakeholders of school organization has helped in increasing consensus to the test and facilitated its acceptance.

On the basis of the results of the report that was sent to their district’s teachers the nine inspectors have started to work with the teachers to develop a reflective action about results and competencies. In the next school year (2013-2014) some courses of continuous learning will start taking into account the test results and especially going into deep the most frequent errors done by the students. In the organization of these courses teachers, inspectors and specialists in the didactics of math will be involved.

If used in a collaborative way an evaluation event may produce positive reaction and be a starting point for a work of innovation. Several teachers appreciated that the report gave them the opportunity to see their class compared with others.

Receiving a feedback in which the performance of their class is compared with the canton and district average scores is less undermining of their representation of their professional self-representation teacher than a feedback

with a comparison between their class performance and the average scores obtained by other specific classes. We opted for this solution in order to avoid teachers to feel the necessity to defend themselves from the evaluation and the feedback they received. The assessment results can in fact become tools for improvement only if people who receive them can integrate them into their self-image.

In spring 2013 a new project of evaluation of mathematics competency in primary school has been started. It concerns other six different dimensions and the test will be administered among in the third class of primary school.

Acknowledgements

The authors thank Urs Moser and Stéphanie Berger of the Institut für Bildungsevaluation of the University of Zurich for their support in the items analysis.

References

Barrier-Ferreira, J. (2008). Producing commodities or educating children? Nurturing the personal growth of students in the face of standardized testing. *The Clearing House, 81*(3), 138-140.

Boaler, J. (2003). When learning no longer matters: Standardized testing and the creation of inequality. *Phi Delta Kappan, 84*(7), 502-506.

Boncori L. (1993). *Teoria e tecniche dei test.* Bollati Boringhieri, Torino.

Buck, S., Ritter, G. W., Jensen, N. C., & Rose, C. P. (2010). Teachers say the most interesting things – An alternative view of testing. *Phi Delta Kappan, 91*(6), 50-54.

Cattell J. McK. (1890). Mental tests and measurements. *Mind.* 15, 373 – 380.

Consorzio PISA. ch (2011). *PISA 2009: Risultati regionali e cantonali*. Berna e Neuchâtel: UFFT/CDPE

De Battisti F., Salini S., Crescentini A. (2006). Statistical Calibration of Psychometric Test. *Statistica & Applicazioni*, Vol. IV, n. 2.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of high stakes testing reforms. *Education & Culture, 23*(1), 55-72.

Penner A. (2003). International gender X item difficulty interacions in mathematics and science achievement tests. *Journal of Educational Psychology*, 95, 650-655

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed.). Boston, MA: Pearson Education, Inc.

Rasch G., (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Woolfolk, A. (2007). *Educational psychology* (10th ed.). Boston, MA: Pearson Education, Inc.