7th International Conference on Higher Education Advances (HEAd'21)
Universitat Politècnica de València, València, 2021
DOI: http://dx.doi.org/10.4995/HEAd21.2021.13046

Verification of the measuring properties and content validity of a computer based MST test for the estimation of mathematics skills in Grade 10

Emanuela Botta

Department of developmental psychology and educational research, Sapienza University of Rome, Italy.

Abstract

The research is aimed at the construction of a multi-level adaptive test (MST), for the evaluation of the mathematical skills of Italian students of Grade 10, and was carried out in collaboration with Invalsi for a PhD study of "La Sapienza" University of Rome. The research started from the definition of the construct to be measured, taking into account both national and international references. A specific item bank was then built. The test was administered to a sample of 4132 students. The experiment confirmed the advantages of an MST model. Interesting results emerged by comparing the adaptive part of the main paths with a linear tests consisting of the same number of items and administered to a sample of pre-test students and comparing the MST test with a simulated linear test, built on the same item bank and with the same numer of item of MST test.

Keywords: *Adaptive test; multistage test; mathematical skills.*

1. Introduction

The introduction of computer-based administration has introduced many changes in large-scale educational assessments, but not all computer-based tests are created equal. The most common are computer-based linear tests, which generally administer a predefined set of items, and variable form tests, in which the computational and interactivity potential offered by the computer is used to administer a set of items that is determined when the test is carried out. CAT, Computer Adaptive Testing, and MST, Multistage Testing, fall into this last category. Many studies state that the adoption of adaptive tests would allow to overcome some of the limitations of the linear ones (Weiss, 1985; Weiss and Kingsbury, 1984; Hambleton, Swaminathan and Rogers, 1991).

The simple transition from paper to computer-based support in linear test maintains the problem that the evaluation carried out is very accurate for average levels of ability but not so much for the extreme levels, with a waste of time and resources in the administration of items to students for whom they are too easy or too difficult to have psychometric value. Only items whose level of difficulty is adequate can significantly contribute to the estimation of the student's ability. In a well-designed adaptive model, it is possible to make the student support mainly items that have a level of difficulty appropriate to his skill level.

Since Grade 10 students generally have a wide range of skills, the test must measure progress across a broad spectrum of outcomes and must be sensitive to small, but significant, academic progress. These two needs are best met by adaptive tests rather than by a linear test because the former adjust the difficulty of the test based on student achievement during the test itself (Hambleton, Zaal & Pieters, 1991; Sands, Waters & McBride, 1997; Sireci , 2004; Wainer, 2000).

2. CAT and MST

In CATs the adaptation of the test to the student's ability takes place after the administration of each item, often starting from a skill level estimated on the basis of other parameters. The design of a multistage test is characterized by the fact that each level is represented by a set of items, called module or testlet, of predefined difficulty; the adaptation of the test to the student's ability therefore takes place on the basis of the cumulative performance on a set of items rather than on the result obtained in each individual item, as in the CAT. In both cases, if the student is doing well, he will be given a more difficult item or set of items, vice versa, an easier item or set of items. A widespread risk of administering CAT tests is that they are not balanced in terms of content. To give an example with the estimation of mathematics ability, it could happen that one student was administered mainly items coming from a single domain, for example arithmetic, and another item coming mostly from another domain, such as geometry. Multi-level tests offer the advantage of greater control over the assembly of forms and the validity of content compared to the item-by-item adaptive tests (*Hambleton*, Swaminathan e Rogers 1991; Hendrickson 2007; Vispoel 1998; Wainer, Lewis, Kaplan e Braswell 1990; Yen, 1993).

3. Framework

There are elements which characterize mathematical competence[1] that are not detectable by means of a standardized test. It was necessary to proceed with the identification of the aspects that could be effectively measured by a standardized computer-based test. It is possible to verify the ability to mathematically formulate a problematic situation or to understand the validity of argumentation but the inclination to use mathematical models of thought is not measurable. So in this test we assume that mathematical competencies are: knowledge and skills and the capacity to apply them to problem solving and to understanding and producing argumentation and reasoning.

The items belong to 4 different content categories, Algebra and arithmetic (AA), Uncertainty and Data (UD), Relationships and functions (RF), Space and shape (SS), and three cognitive domain Knowledge, Problem Solving and Reasoning, then articulated in general and specific learning objectives. The cognitive domains constitute a grouping of goals (learning outcomes), based on the idea that the mathematical activities essentially refer to either reasoning or solving problems and that these two dimensions are not independent of each other and require knowledge of concepts, formal language and procedures to be implemented. The semiotic dimension of representation is considered cross-cutting to the others and takes on different aspects in each of them.

4. The item bank

A specific item bank was built, by carrying out two distinct pre-test phases anchored together. Sampling of students, distributed throughout Italy and in various types of schools, was always carried out with random assignment of forms to students. In the first phase, 18 test forms (460 items) anchored together were administered to a sample of 4672 students. In the second phase, 11 test forms (403 items) were administered to a sample of 5797 students. These tests were anchored to each other and to those of the previous phase. In both phases of the bank construction, the items were selected by calibrating the difficulty of the items according to

[1] For the definition of the mathematics framework, the definition of mathematical competence adopted in the European framework of key competences and those currently present in Italian legislation were taken into consideration. The main reference was the INVALSI Framework, but we also compared ourselves with the main international references: the frameworks for mathematics of OECD PISA and NAEP and the Cambridge IGCSE Syllabus.

the 1-parameter Rasch model, verifying the unidimensionality of the construct with a specific EFA (Exploratory Factor Analysis), and considering: the format of the items, the proportion of correct answers (p > 0,10), discrimination of items (R > 0,20), defined as the biserial point correlation of one item with all others of the same test form, the main fit indices, such as the standardized residue (zResid not significant), Infit index and Outfit index, between 0,8 and 1,2. In each phase, for the calibration of the item bank, the method of concurrent calibration of all test forms was used since it allows to place the estimates of the item parameters and the ability estimates on the same scale without an additional linking procedure. Two factor analysis were carried out for each form: one before the calibration of the bank, to identify the items that did not have a good loading with the factor, and one after the calibration to verify that the unifactorial solution was acceptable. The analysis carried out confirms the hypothesis that the unifactorial solution is correct. The extraction of additional factors has not shown convincing solutions. To obtain a single bank, calibrated on a single scale, the two banks were linked using the Anchor-Test Design with the Robust Mean and Sigma Method (Hambleton, Swaminathan, Rogers, 1991 and Stocking & Lord, 1983), that is usually used to develop a common metric in the Item Response Theory. At the end of the linking process, the estimated values of b for all items are placed on the same scale and it was possible to perform a recalibration of the ability estimate of all students, to fixed values of b. The item bank was made up of 497 items, all on the same scale. Figure 1 illustrates the distribution of items in relation to the difficulty parameter b.

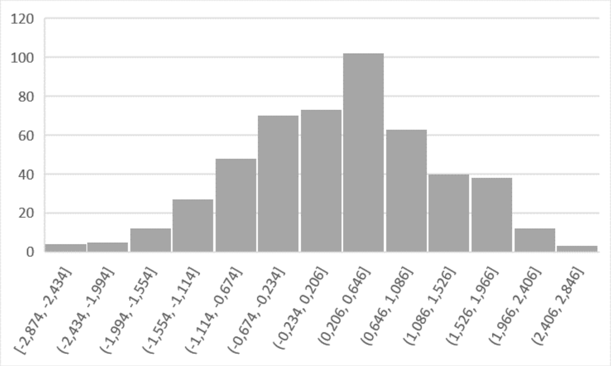

Figure 1. Distribution of items in relation to the difficulty parameter b.

5. The MST test: developing the MST 1-3-3 model

In order to develop an MST 1-3-3 model, like the one in Figure 2, three intervals have been identified along the continuity of the ability, with the central interval placed on the mean ability of the sample. Each range has an amplitude equal to one standard deviation. For Stage 1, a routing module (16 items) was designed. For Stage 2 and 3, two modules were designed for each skill interval, consisting of 18 and 12 items. All modules focused on the average ability of the sample in the reference range (E, easy, M, medium, H hard). The selection of the items took place with an optimization process with the constraint of maximizing the information function I(θ) and balancing the modules in relation to the framework.

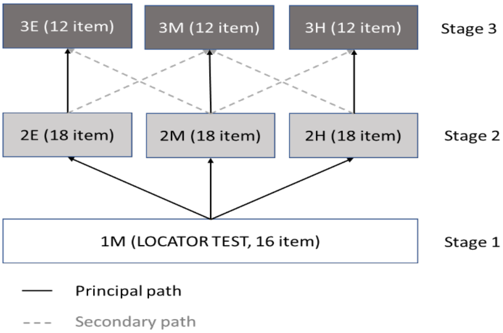

Figure 2. MST 1-3-3 model.

For the definition of the routing rules, the cut-off values of the ability, that identify the threshold for accessing the next module at the end of each module, were identified intersecting the information functions of the modules at stage 2 and 3 and converted into true score values (Luecht, Brumfield, Breithaupt, 2010). The test was administered to a sample of 4132 students equally distributed throughout the country and by field of study. The selection of the students took place with a two-stage sampling.

5.1. Content validity

During the test assembly it was possible to set and comply with stringent constraints covering the framework which guaranteed the validity of the test content. For each content area there are 26 or 27 items in the complete MST test. The distribution of items in each path is balanced with respect to the content areas as you can see in Table 1.

Table 1. Number of items for each path.

	Path	UD	AA	RF	SS
Path 1	1M+2E+3E	12	12	11	11
Path 2	1M+2E+3M	12	12	11	11
Path 3	1M+2M+3E	12	11	12	11
Path 4	1M+2M+3M	12	11	12	11
Path 5	1M+2M+3H	12	11	12	11
Path 6	1M+2H+3M	11	12	11	12
Path 7	1M+2H+3H	11	12	11	12

6. The comparison between the MST test and the linear tests

In order to verify, in terms of measurement accuracy, the actual improvement of the MST test compared to a non-adaptive test centered on the population average, it was decided to make two comparisons: the complete MST test with a linear test built from the same item bank and the adaptive part of the MST with one of the linear tests used to perform the pretest. To do this, it was necessary to choose a linear test with exactly the same number of items as the MST test, 46 items for the complete test and 30 items for the adaptive part of the MST test, because the information function is influenced by the number of items. The comparison is feasible because the items of the two tests are on the same scale, so difficulty and ability are directly comparable.

6.1. The adaptive part of the MST with one of the linear tests used to perform the pretest

Among those available, the linear test that showed the best psychometric characteristics was chosen. For this comparison, the estimation of the students' ability was used, carried out using only the items of the adaptive stages, 2 and 3. This was done to have comparable tests by number of items and to estimate the measurement capacity of the adaptive part of the test.There were two types of tests: a direct comparison of information functions and a t-test for independent samples on the average number of correct answers and on the average value of the SE in the estimate of the ability. For each module only those students who fell into the intervals into which the continuous of the ability in the construction phase of the test was divided, were considered. From the t-test it can be seen that the SE in estimating the ability of the MST test is always significantly lower than that of the linear test. It is also observed that in Path 1 the students respond on average to 6.950 items more than in the linear test, in Path 4 on average to 2.819 items more and in Path 7 on average to 2.441 items less. The MST test is therefore able to offer students with low abilities an adequate number of items to which they are able to respond and those with high abilities items that are still challenging and able to investigate within the proximal development zone. Table 2 reports the values of the information function of the main paths of the MST test and of the Linear test of the pretest, with 30 items, at the extremes of the interval, θ = -2.243 and θ = +1.430, in the middle of the interval, θ = -0.406, and around the decision nodes, θ = -1.06 and θ = +0.21.

Table 2. Information function of MST path vs Linear 30.

θ	**Path MST**	**Linear**	**MST – Linear**
-2.243	6.49	2.49	4.00
-1.070	6.63	4.82	1.81
-0.406	7.34	6.02	1.32
+0.210	6.72	6.50	0.22
+1.430	6.71	5.15	1.56

6.2. The complete MST test vs a linear test built from the same item bank

In this case, a direct comparison was made between the information functions of the two tests.

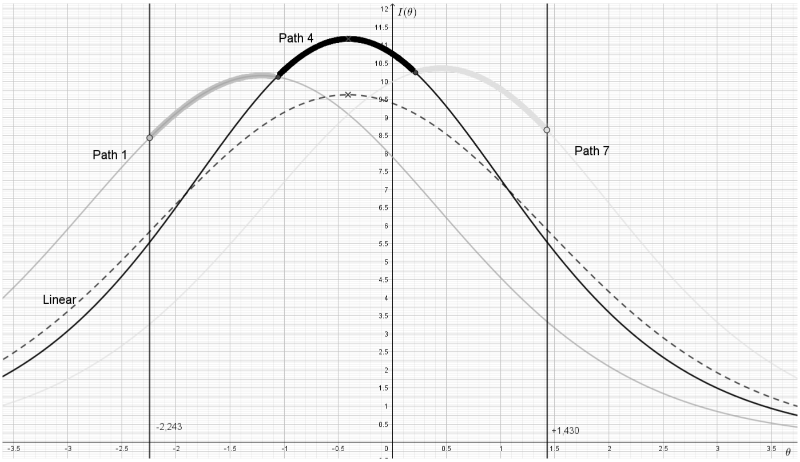

Figure 3. MST vs Linear 46.

Table 3 reports the values of the information function of the main paths of the MST test and of the Linear 46 test, at the extremes of the interval, θ = -2.243 and θ = +1.430, at the center of the interval, θ = -0.406, and around at the decision nodes, θ = -1.06 and θ = +0.21.

Table 3. Information function of MST path vs Linear 46.

θ	Path MST	Linear 46	MST – Linear 46
-2.243	8.43	5.54	2.89
-1.060	10.12	9.04	1.08
-0.406	11.12	9.63	1.49
+0.210	10.24	9.10	1.14
+1.430	8.64	5.54	3.1

7. Conclusions

The article shows that the limitations of linear and adaptive tests are overcome. The test turns out to be more informative than a linear test built on the same item bank and allows for more reliable estimates of student ability within a wide range of the continuum, improving estimates particularly for students at the extremes of the range. At the same time, the test succeeds in offering students with low ability a fair number of items that they are actually able to answer and students with very high ability items that are still challenging.

References

Weiss, D. J., e Kingsbury, G. G. (1984). *Application of computerized adaptive testing to educational problems*. Journal of Educational Measurement, 21(4), pp. 361-375.

Weiss, D. J. (1985). *Adaptive testing by computer*. Journal of consulting and cli-nical psychology, 53(6), p. 774.

Hambleton R. K., Swaminathan H., Rogers H. J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Inc. 1991, London.

Hambleton, R. K., Zaal, J. N., e Pieters, P. (1991). *Computerized adaptive testing: Theory, applications, and standards*. In R. K. Hambleton e J. N. Zaal (Eds.), *Advances in educational and psychological testing*, Norwell, MA: Kluwer, pp. 341–366.

Sands, W. A., Waters, B. K., e McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Sireci, S. G. (2004). *Computerized-adaptive testing: An introduction*. In J. Wall e G. Walz (Eds). *Measuring up: Assessment issues for teachers, counselors, and administrators*, Greensboro, NC: CAPS Press, pp. 685-694.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Stocking, M. L., & Lord, F. M. (1983). *Developing a common metric in item re-sponse theory*. Applied psychological measurement, 7(2), 201-210.

Luecht, R., Brumfield, T., e Breithaupt, K. (2006). *A testlet assembly design for adaptive multistage tests*. Applied Measurement in Education, 19(3), pp. 189-202.

Hendrickson, A. (2007). *An NCME instructional module on multistage testing*. Educational Measurement: Issues and Practice 26, pp. 44-52.

Vispoel, W.P. (1998). *Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests*. Journal of Educational Measurement 35, pp. 328-345.

Wainer, H., C. Lewis, B. Kaplan, e J. Braswell. (1990). *An adaptive algebra test: A testlet-based, hierarchically structured test with validity-based scoring*. Technical Report 90-92. Princeton. NJ: Educational Testing Service.

Yen, W. M. 1993. *Scaling performance assessments: Strategies for managing local item dependence*. Journal of Educational Measurement 30, pp. 187 -214.