

Large-Scale Assessment as a Tool for Monitoring Learning and Teaching: The Case of Flanders, Belgium

Erik De Corte*, Rianne Janssen**, & Lieven Verschaffel*

* Center for Instructional Psychology and Technology (CIP&T)

** Center for Educational Effectiveness and Evaluation

University of Leuven, Belgium

Abstract

Traditional tests for large-scale assessment of mathematics learning have been criticized for several reasons, such as their mismatch between the vision of mathematical competence and the content covered by the test, and their failure to provide relevant information for guiding further learning and instruction. To achieve that large-scale assessments can function as tools for monitoring and improving learning and teaching, one has to move away from the rationale, the constraints, and the practices of traditional tests. As an illustration this paper presents an alternative approach to large-scale assessment of elementary school mathematics developed in Flanders, Belgium. Using models of item response theory, 14 measurement scales were constructed, each representing a cluster of curriculum standards and covering as a whole the mathematics curriculum relating to numbers, measurement and geometry. A representative sample of 5,763 sixth-graders (12-year-olds) belonging to 184 schools participated in the study. Based on expert judgments a cut-off score was set that determines the minimum level that students must achieve on each scale to master the standards. Overall, the more innovative curriculum standards were mastered less well than the more traditional ones. Few gender differences in performance were observed. The advantages of this approach and its further development are discussed.

Introduction

Assessment is concerned with the design, construction, and use of instruments for determining how powerful learning environments are in facilitating in students the acquisition of the different aspects of competence in a domain, e.g., mathematics. Assessments of learning can either be internal or external. Internal assessments are organized by the teacher in the classroom, formally or more informally; to the contrary, external, usually large-scale assessments come from outside, organized at the district, state, national, or even international level using standardized tests or surveys. As argued by the National Research Council (2001) in the US, assessments in both the classroom or a large-scale context can be set up for three broad purposes: to assist learning and teaching, to measure achievement of individual pupils, or to evaluate school programs. I like to argue that a major purpose should be to use assessment *for* learning which means that it should provide useful information for students and teachers in view of fostering and optimizing further learning. Sloane and Kelly (2003) contrast assessment *for* learning or formative assessment with assessment *of* learning, the goal of the latter being to determine what students can, and whether they attain a certain achievement or proficiency level. In this paper I will focus on large-scale assessment of mathematics education in Flemish primary school.

Large-scale assessment of learning: A critical discussion

The massive use of standardized tests in education has always been more customary in the United States as compared, for instance, to Europe. But especially since the beginning of the 1990s the traditional tests have been criticized.

Analyses of widely used standardized tests show that there is a mismatch between the new vision of competence in different domains, on the one hand, and the content covered by those tests, on the other hand. Due to the excessive use of multiple-choice item format, the tests focus on the assessment of memorized facts, rote knowledge, and lower-level procedural skills. On the other hand, they do not sufficiently yield relevant and useful information on pupils' abilities in problem solving, in modeling complex situations, in communicating ideas, and in other higher-order thinking skills. A related criticism points to the one-sided orientation of the tests toward the products of pupils' mathematics work, and the neglect of the processes underlying those products.

An important consequence of this state-of-the-art is that assessment often has a negative impact on the implemented curriculum, the classroom climate, and instructional practices, dubbed the WYTIWYG ("What You Test Is What You Get") principle (Bell, Burkhardt, & Swan, 1992). Indeed, the tests as characterized above convey an implicit message to students and teachers that only facts, standard

procedures, and lower-level skills are important and valued in mathematics education. As a result teachers tend to ‘teach to the test’, i.e. they adapt and narrow their instruction in the sense that they give a disproportionate amount of attention to the teaching of the low-level knowledge and skills addressed by the test at the expense of teaching for understanding, reasoning, and problem solving.

An additional major disadvantage of the majority of traditional evaluation instruments is that they are disconnected from learning and teaching. Indeed, also due to their static and product-oriented nature most achievement measures do not provide feedback about students’ understanding of basic concepts, nor about their thinking and problem-solving processes. Hence, they fail to provide relevant information that is helpful for students and teachers in view of guiding further learning and instruction.

Apart from the previous intrinsic criticisms on traditional standardized achievement tests, a major issue of debate is their accountability use as high-stake tests, i.e. their mandatory administration for collecting data on the attainment of students as a basis for highly consequential decisions about students (e.g., graduation), teachers (e.g., financial rewards), and schools and school districts (e.g., accreditation). According to the No Child Left Behind Act in the US this accountability use should result in the progressive acquisition by all students of a proficiency level in reading and mathematics. However, a crucial question is whether current testing programs really foster and improve learning and instruction; and, there are serious doubts in this regard. In a study by Amrein and Berliner (2002) involving 18 US states it was shown that there is no compelling evidence at all for increased student learning, the intended outcome of those states high-stake testing programs. Moreover, there are many reports of unintended but unfavorable consequences, such as increased drop-out rates, negative impact on minority and special education children, cheating on examinations by teachers and students, teachers leaving the profession, etc. In addition students tend to focus on learning for the test at the expense of the broader scope of the standards.

The Flemish approach to large-scale assessment of mathematics at the primary school

To achieve that large-scale assessments do indeed foster and improve student learning, one will have to move away from the rationale, the constraints, and the practices of high-stake testing programs. As one example I will briefly review here an alternative approach to large-scale testing developed in the Flemish part of Belgium (for a more detailed discussion see Janssen, De Corte, Verschaffel, Knoors, & Colémont, 2002).

In a project commissioned by the Department of Education of the Flemish Ministry, we developed an instrument for the national assessment of the new standards of the entire new mathematics curriculum. These standards represent the basic competencies that students should master at the end of the primary school (which consists of 6 grades starting at the age of 6). The instrument was used to obtain a first, large-scale baseline assessment of the attainment of those new curriculum standards. The aim was thus not to evaluate individual children or schools as a basis for taking high-stake decisions, but to get an overall picture of the state-of-the-art of achievement in mathematics across Flanders at the end of primary education. The instrument consists of 14 measurement scales, each representing a cluster of standards and covering as a whole the entire mathematics curriculum relating to numbers, measurement, and geometry.

Using a stratified sampling design, a representative sample of 5763 sixth-graders (12-year-olds) belonging to 184 schools participated in the investigation. Taking into account the aim of the assessment it was not necessary to have individual scores of all students, and a population sampling approach could be used

“whereby different students take different portions of a much larger assessment, and the results are combined to obtain an aggregate picture of student achievement” (Chudowsky & Pellegrino, 2003, p. 80).

This approach also allows to really cover the total breadth of the curriculum standards. More specifically, the instrument involved 10 booklets, each containing about 40 items belonging to two or three of the 14 measurement scales. To get booklets that were somewhat varied the scales in each booklet represented distinct mathematical contents (for instance, the items in booklet 2 related to ‘percentages’ and ‘problem solving’). Each booklet was administered to a sample of over 500 sixth graders. Four different item formats were used: short-answer (67%), short-answer with several subquestions (14%), multiple-choice (11%), and product and process questions (8%). Especially the

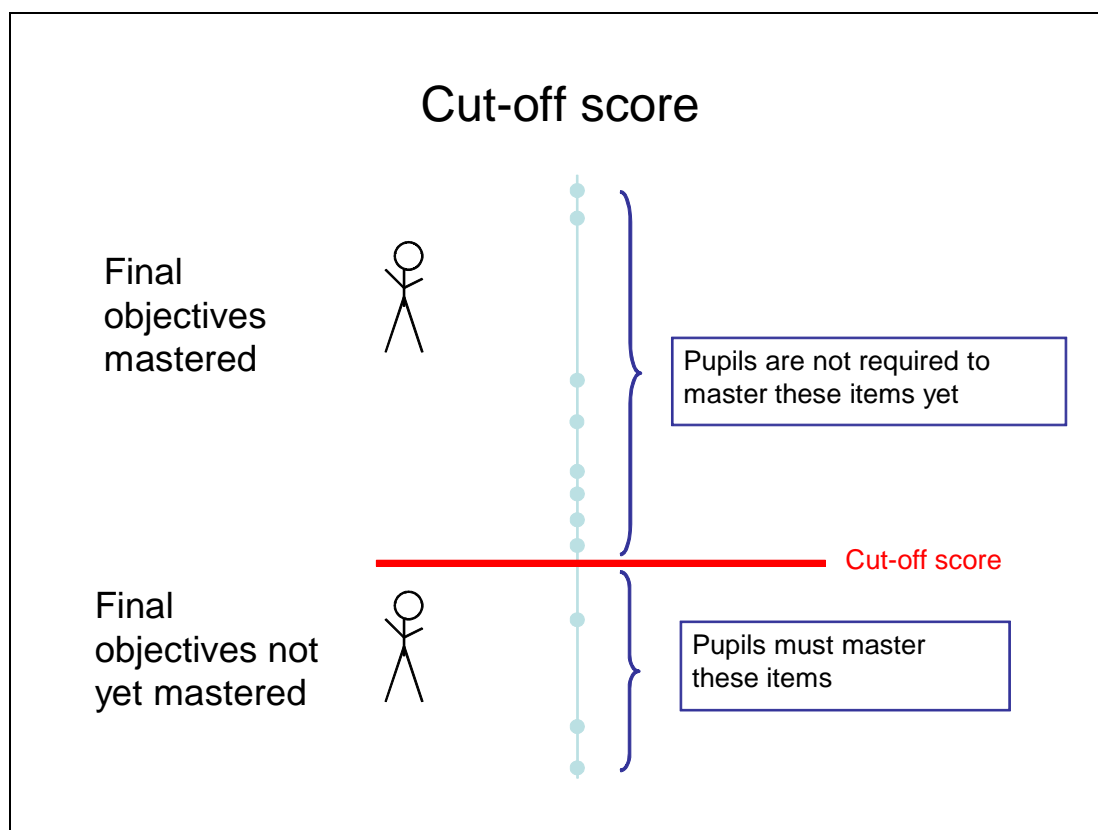


Figure 3. Cut-off score showing the divide of the items as well as the students. The next step consisted in determining on the measurement scales the minimum level that students must achieve in terms of the test items. Indeed, the standards describe that basic competencies in general terms, but because for each standard items of very different difficulty levels can be developed, there is a need to set a cut-off score that defines the minimum level of competence for each scale. This was done by consulting a group of expert judges who were asked to set the cut-off score for each scale based on a careful analysis of the content of the items. The cut-off score distinguishes the items the students must master well to attain the standards or basic competencies and the items that go beyond the minimum level (see Figure 3).

The results of this assessment shown in Table 1 can be summarized as follows. Scales about declarative knowledge and those involving lower-order mathematical procedures were mastered best. The scales relating to more complex procedures (e.g., calculating percentages; calculating perimeter, area, volume), and those that address higher-order thinking skills (problem solving; estimation and approximation) were not so well mastered. The latter finding is not so surprising as those scales relate to standards that are relatively new in the Flemish mathematics curriculum. It is also interesting to mention that few gender difference in performance were observed.

It is the intention of the Department of Education of the Flemish Ministry to organize such a large-scale assessment of mathematics education periodically in the future. The next assessment will take place in May 2009. The advantages and the potential of this approach to large-scale assessment are obvious. First, because this assessment covers the entire curriculum, its findings are a good starting point for continued discussion and reflection on the standards in and among all education stakeholders (policy makers, teachers, supervisors and educational counsellors, parents, pupils). Second, due to the breadth of such an assessment approach, it uncovers those (sets of) standards that are insufficiently mastered. In doing so the assessment provides relevant feedback to practitioners (curriculum makers, teachers, counsellors) by identifying those aspect of the curriculum that need special attention

in learning and instruction; and researchers could focus intervention research on those weaknesses in pupils' competence. Third, due to the alignment of the assessment and the curriculum the often heard complaint about 'teaching and learning to the test' can largely be avoided, especially if appropriate counselling and follow-up care is provided after the results are published. Moreover, because the Ministry does not at all intend to use the results for the evaluation of individual teachers or schools, and because scores of individual children, classes, or schools are not published, the negative consequences of high-stake testing referred to above are also avoided.

Table 1. Overview of the assessment results for the 14 measurement scales

Measures in meaningful context	88%
Units of measure	88%
Geometry: concepts	87%
Space and spatial orientation	86%
Number values and equivalence	86%
Ratios	74%
Reference points	72%
Problem solving: measurement and geometry	68%
Problem solving: numbers and operations	68%
Fractions and decimal numbers	64%
Rounding off, estimation, approximation	63%
Meaningful reductions	56%
Perimeter, surface area, volume	53%
Calculating percentages	42%

References

- Amrein, A.L., & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning. *Educational Policy Analysis Archives*, 10 (18). (<http://epaa.asu.edu/epaa/v10n18/>).
- Bell, A, Burkhardt, H., & Swan, M. (1992). Balanced assessment of mathematical performance. In R. Lesh & S.J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 119-144). Washington, DC: American Association for the Advancement of Science.
- Chudowsky, N., & Pellegrino, J.W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice*, 42, 75-83.
- Janssen, R., De Corte, E., Verschaffel, L., Knoors, E., & Colémont, A. (2002). National assessment of new standards for mathematics in elementary education in Flanders. *Educational Research and Evaluation*, 8, 197-225.
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Sloane, F.C., & Kelly, A.E. (2003). Issues in high-stakes testing programs. *Theory into Practice*, 42, 12-17.