Standards, Standard setting and Accountability in the Netherlands.

37th IAEA Annual Conference
Manila, 23-28 October 2011,

Frans Kleintjes, Senior Research Scientist,

Cito, Arnhem The Netherlands

Abstract

Declining PISA results and disappointing results on Computer Adaptive entrance tests for teacher training result in a national initiative in the Netherlands to raise the level of student's performance on Dutch language and mathematics at different levels of education. For this purpose, Cito has been asked to develop tests corresponding with consecutive reference levels at different stages in a student career: end of primary education, end of basic education, end of secondary education and end of senior secondary education. At each stage basic competence levels are determined. Some of these levels correspond with national examination levels. Furthermore, Cito is involved in the development of fully adaptive exams. In vocational education starting in 2014, based on Cito's experience with the teacher training entrance tests, adaptive exams will be available.
In this presentation we will discuss different views on standards and standard setting procedures in relation to schools accountability and high stakes testing.

Introduction

The Dutch decentralized education policy, gives schools freedom and makes them responsible for the quality of the education they provide. Because schools are held responsible for the quality of education, they have to implement an active policy of quality control, on both student level and school level.

For a long time educational assessment during a students' school career has been the field of the school, and its teachers. Teachers and school constructed, used and interpreted their own tests. This has gradually changed over the years from only test developed by teacher and school to the use of more standardized testing and testing in monitoring systems. More and more the teaching community has integrated standardized testing as part of the assessment at schools. This does not mean that tests developed by the school have been taken over by standardized testing. Besides standardized testing tests developed by schools are still needed to evaluate the teaching process.

The use of standardized tests, for instance in monitoring systems or at final examinations, creates new possibilities in comparing schools, groups and teachers, simply because data are available and therefore in accounting for achievements. This is however not as simple and easy as some policymakers like to believe. Education is a complex process as is the interpretation of educational assessments results in view of accountability. Even if tests are developed especially for accountability, valid testing is complicated

Concern about performance level on basic education.

Declining PISA results and disappointing results on Computer Adaptive entrance tests for teacher training result in a national initiative in the Netherlands to raise the level of student's performance on Dutch language and mathematics at different levels of education. For this purpose, Cito was asked to develop tests corresponding with consecutive reference levels at different stages in a student career: end of primary education, end of basic education, end of secondary education and end of senior secondary education. At each stage basic competence levels are determined. Some of these levels correspond with national examination levels.
A similar concern was expressed in the USA in the early 1980s (Koretz and Hamilton 2006 p 534). In the USA this resulted in heavy debates: should Standards Reference testing only involve expectations on minimum competence or should they involve standards for high expectations as well. This standard movement created a rapid move towards reporting of performance in terms of performance standards rather than in terms of norm-referenced or other conventional scales. The primary motivation according to Koretz and Hamilton for this shift appears to have been a desire to compare performance to expectation –how much students should know- rather than a distribution of current performance, which many reformers considered unacceptably low.

Do we expect to move into a similar debate in the Netherlands soon? One of the largest advantages of describing minimum competencies in Dutch language and mathematics on all levels in education is that it is (one of) the first time(s) that requirements have been developed over a student's career. So not isolated at different periods in a student's life but a coherent set over time. From this framework it can be expected that a more detailed description per year will be required soon, and we can expect that the debate on inclusion of higher order skills in view of validity of the tests will start soon. As indicated by Hambleton and Pitoniak, there is a need for new ideas and more research in this area. Probably the most important topic for research today is the vertical alignment of performance standards across grades within a

subject, and across subjects at grade level. Another important topic, because it is the final step prior to performance standards becoming policy, concerns the best ways to present the results of a standard setting study to a policy board.

We will indicate how Cito attempts to take into account above considerations and experiences in the tests and exams that are developed. We will restrict ourselves to some selected educational assessments that are produced by Cito or in assessment instruments that are constructed under Cito's responsibility. Cito has long developed standardized tests of which aggregated test results on group or school level have been used for school self-evaluation but not for accountability in a sense of comparing and awarding schools.
However, in The Netherlands the Inspectorate is the institute that autonomously uses these standardized test results for accountability. Standardized testing has advantages, but there are threats when the tests are used for accountability. A plea is made to ensure that test results will be made useful for all stakeholders. Therefore test developers must (start to) take as many considerations into account as possible.

Tests for accountability

Accounting for results means that there must be expectations against which can be reported. If results of educational tests are available they will be used for accounting if they are comparable. The more comparable and the simpler the reporting the more likely they will be used for accounting purposes. Accounting therefore has an opportunistic character: If comparable data are available they will be used, whether the tests have been developed for accounting purpose or not. Koretz and Hamilton in their chapter in Educational Measurement notice the changes in educational testing practice over time into testing for accountability and mention the complexness of testing for accountability. They point at desirable and undesirable steps that educators take to prepare students for tests: teaching more, working harder, working more efficiently, reallocation, alignment, coaching and cheating. Because of the pressures for test based accountability, the potential for corruption or inflation of test scores must now be a central concern in the evaluation of validity.
Keeping all of this in mind we will focus first on how to arrive at expectations against which can be reported.

Standards

Whenever educational assessment is used to categorize individuals, performance standards must be established along a score range (Hambleton and Pitoniak, 2006). Many different terms are used in the measurement literature to refer to performance standards They may be referred to as" 'passing scores', 'cut scores', 'cutoff score', 'performance levels', 'achievement levels' , 'mastery levels', 'proficiency levels', 'tresholds' and 'standards'. Examinees may be classified as 'pass' or 'fail, or may be into a greater number of ordered performance categories with labels such as 'below basic' , 'basic', 'proficient', and 'advanced'. Performance categories are the intervals between the performance standards on the score reporting scale. In practice detailed descriptions of the knowledge and skills of candidates located in these performance categories are developed and used to communicate test results (Hambleton, 2001)

The importance of setting proper and valid performance standards has been highlighted in recent years as the use of assessment for accountability purposes in education has increased. Numerous methods are available for setting performance standards. Most used are judgmental methods such as criterion referenced tests and the more familiar norm referenced approaches. Norm-referenced methods determine pre-specified percentages of examinees to pass or fail. In contrast to norm-referenced methods, criterion-referenced methods use content standards to outline the knowledge, skills, and abilities as the basis of judgments. Norm-referenced passing scores are not suitable for high school graduation, licensure, or certification tests because scores are interpreted with respect to performing better or worse than others, rather than with respect to the level of competence of a specific test taker.

The word 'standards' can be used both in conjunction with the content and skills candidates are viewed as needing to attain and the scores they need to obtain in order to demonstrate the relevant knowledge and skills. In the context of educational assessment, Hambleton and Pitoniak make therefore a distinction between content standards and performance standards, since confusion about these concepts often arise among policymakers, educators, and the public. The establishment of content standards often is the domain of policy makers and educators. Content standards are reflected in the curricula, and specify what examinees are expected to know and to be able to do. Content standards provide direction to instructors on what they need to teach. Performance standards, in contrast define the levels of test performance examinees are expected to attain in relation to the content standards. Performance standards may thus be viewed as an operationalization of the content standards

to the test or assessment that has been constructed to measure the content standards. In educational testing the number of performance standards often varies from one to as many as 10 to distinguish passing and failing as is the case in the Dutch marking system.

Classifying standard setting methods

The long used classification into criterion and norm referenced or test-centered and examinee centered methods, has been extended into four categories (Hambleton, Jaeger, et al, 2000) Review of test items and scoring rubrics, review of candidates, looking at candidate work and panelists review of score profiles

Hambleton and Pitoniak classify available standard setting methods along these four categories. They also describe compromise methods that take into account both absolute and relative standards. For a listing and description of the methods we refer to Hambleton and Pitoniak (2006). Cizek (2006) describes ten methods that are in use most. We list them briefly below.

Methods of standard setting.

Kleintjes and & Moelands review a selection of tests developed by Cito and discuss how standards are set and have been incorporated. They indicate the development of use of test results for accountability for these tests.

In this section a brief overview of standard setting methods is presented. For a more extensive overview we refer to the chapter on standard setting by Cizek in the Handbook of test Development (Downing and Haladyna eds, 2006).

As Cizek indicates in the introduction to his chapter there can be no single set of procedures for determining cut scores for all test and all purposes, nor can there be any single set of procedures for establishing their defensibility. This will also become clear by illustrating the different methods used by Cito. Although they depart from one of the method listed, they divert more or less form the method when they are applied in real test situations. Each situation requires adaptation of a method to local use and acceptance.

There is continuously growing list of methodological options for setting standards. Cizek limits the list to the methods that are most frequently used.

The Angoff method:

Keeping a minimally performing person in mind, one should go through the test item by item and decide whether such a person could answer correctly each items under consideration. A

score of one is given for each item that is answered correctly by that hypothetical person. The sum of the itemscores will equal the raw score earned by the minimally performing person. A slight variation has become a typical application: Indicate for each item the probability that the minimally acceptable person would answer the item correctly. The sum of these probabilities would than represent the minimally acceptable score.

Angoff variations

- The Yes/No method
  Substitutes Yes/No instead of the 1/0 in the original Angoff procedure.
- The extended Angoff method
  Uses a mix of selected response and constructed response

The Nedelsky method,

Using possible eliminations of alternatives by minimal competent candidates, the reciprocal of the remaining alternatives is used as minimal required score. The sum of these values can be directly translated into a passing score..

The Bookmark method.

This procedure is based on placing a bookmark or bookmarks in a specially prepared booklet in which the items and tasks are ordered according to difficulty, from easiest to hardest. The bookmark is placed where an examinee on the borderline will answer the item correctly, with say 67 percent probability. The percentage may differ depending on the nature of the research. The bookmark corresponds to the ability required by a minimal competent candidate. As this value is on the ability scale this can be translated to the score distribution of any subset of items that are on the scale. Note that is fairly easy to establish more than one cut point.

The Contrasting group

Judges are asked to make direct judgments about the real status of examinees with information about their actual performance on an examination. Information about the candidates with respect to the characteristic to be assessed is required to from two groups: the true non-masters and the true masters. The score distribution of the two groups are plotted and analyzed. The cut score is established at the intersection of the two distributions.

The Borderline method

Judges, familiar with the specific knowledge, skills, and abilities of individual examinees who are subject to the examination, use this special knowledge to try to describe the borderline between mastery and non- mastery. Without knowledge about the examinees performance participants will identify candidates that are on the borderline of acceptable and unacceptable competence. Often the median of the scores of these borderline examinees is identified as the cut score.

The Body of Work method

Examinee works (test performances) are assigned to categories like master and non-master. Examinee works are scores prior to the standard setting. The works are selected to span the range of obtainable scores. Participants, again without knowledge about the scores assigned to the work samples, than rate each work holistically and classify it into one of the required categories. Cut scores are obtained next by using the intersections of the score distributions in each category.

Methods for adjusting Cut Scores

Assuming that the psychometric procedures for standard setting were carried out with fidelity any adjustments of standards are necessarily based more on policy considerations than on technical bases. Two methods for striking a compromise between 'absolute methods' and norm-references approaches are mentioned below. Both method ask participants explicitly to state the pass and fail rates the they believe to be reflective of the 'true' proportions in the sample of examinees and tolerable from political, economic or other perspectives.

The Beuk method

Each judge is asked two questions:

1. What should be the minimum level of knowledge to pass an examination? and
2. What passing rate should be expected?

The ratio of the standard deviations of the percentage correct and the passing rate is used in the calculation the obtain the adjusted or compromised percent correct, that is the cut score.

The Hofstee method

Each participant is now asked four questions:

1. What is the lowest cutoff score that would be acceptable, even if every student attained that score on the first testing?
2. What is the lowest acceptable cutoff score, even if no student attained that score on the first testing?

3. What is the maximum tolerable failure rate?

4. What is the minimum acceptable failure rate?

Mean values across participants are calculated in respect to each question, to obtain a line through the two points (minimum acceptable failure rate, lowest acceptable cut off score) and (maximum tolerable failure rate, lowest cutoff score even if every student attained that score). The point on the line represent acceptable percentages fails in combination with percent correct required. The intersection of this line and the observed score distribution yields a compromised cut score.

Change in use of Cito tests

Some important tests in The Netherlands that are produced by Cito have undergone changes in use towards accountability gradually in recent years. Central examinations at the end of secondary education, and a compulsory external test at the end of primary education have been the main tools in monitoring and controlling the quality of education in the Netherlands. Schools used the results mainly in school self-evaluation.

Rather than being used by schools for school self-evaluation, the inspectorate nowadays uses the results on the standardized tests to measure the output, and holds schools accountable for the results. But not only evaluation at the end of a student's career is used for accountability nowadays. Although for a long time the main focus of evaluating outcomes used to be at the end of educational trajectories, standardized testing in a framework of student monitoring could provide valuable additional information, during the educational career.

However, not only the development and use of tests has changed over years. Some developments at Cito have taken place as well. Cito was established in 1968, and used to be a governmental department. In 1999 Cito was privatized (Roorda, 2010). The main impact of this change has been a shift towards delivery on demand of schools instead of delivering tests to schools that had been products on behalf of the ministry. Nowadays, more attention is paid to the needs and wishes of schools. This places test institutes like Cito in a kind double role when developing tests for accountability.

Standardised tests are developed based on the wishes of schools to justify their education by measuring students' progress. However, the Dutch inspectorate uses the outcomes of the various standardized tests as one of the components of their evaluation system to evaluate the schools as well. When tests are used for accountability, this will sure have an impact on the behavior of schools against tests. This is the focus of the contribution of Hermans and

Wiegers (2011) in this very session.

National and international accredited institutes

Assessment for accountability requires a high level of expertise in educational assessment of all stakeholders. All parties will have to act according to professional standards, this will requires continuous professional development. We propose to establish a system to accredit institutes and certify professionals within these institutes involved in educational measurement and also to certify providers of data for accountability to ensure a sound, valid and justifiable accountability process.

Establishing such an institute not only at national level but one at international level would enable member countries to even audit their educational systems in a professional way.

References.

Downing, Steven M. and Haladyna, Thomas M. (Eds.) Handbook of Test Development, London 2006.

Hambleton, R.K. and .Pitoniak, M.J. Setting Performance Standards in Educational Measurement 4th edition edited by Brennan, R.L, 2006 NCME, ACE. USA.

Korets, D.M. and Hamilton, L.S., Testing for accountability in K-12. in Educational Measurement 4th edition edited by Brennan, R.L, 2006 NCME, ACE. USA.

Roorda, M. (2010). Educational assessment- a private matter? Key note at the annual conference of the AEA-Europe. November 2010 Oslo, Norway.

Lubbe, M. van der (2007) The End of Primary School Test. Paper presented at the at 33th IAEA Annual Conference, Azerbaijan

Lubbe, M. van der (2007). Pupil Monitoring System (PMS) for Primary Education. Paper presented at the at 33th IAEA Annual Conference, Azerbaijan,

Moelands, H. (2010). Computerized adaptive testing in the Monitoring and Evaluation System for .primary education in The Netherlands. Paper presented at 36th IAEA Annual Conference
Bangkok, Thailand, August 22 – 27, 2010

Van der Schoot, F. (2001). Standaarden voor kerndoelen basisonderwijs. Doctoral dissertation, University of Amsterdam.

Van der Schoot, F. (2009) Cito variation on the bookmark method. Section I in the Reference Supplement to the Manual for Relating language examinatiosn to the Common European Framework of Reference for Languages. Language Policy Division, Strasbourg. www.coe.int/lang.

Verhelst, N.D. & Eggen, Th.J.H.M. (1989). Psychometrische en statistische aspecten van peilingsonderzoek. PPON-rapport nr. 4. Arnhem: Cito.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1993). OPLM: One Parameter Logistic Model. Computer program and manual. Arnhem: Cito.

Verhelst, N.D. & Glas, C.A.W. (1995). The One Parameter Logistic Model. In G.H.Fischer & I.W. Molenaar (Eds.), Rasch Models: Foundations, Recent Developments and Applications (pp. 215-237). New York: Springer-Verlag.

Kuyper, H., Keuning, J., & Zijsling, D. (2010). *Basisrapport eerste meting van COOL$^{5-18}$ in het derde leerjaar van het voortgezet onderwijs*. Groningen/Arnhem: GION/Cito.

Zijsling, D., Keuning, J., Kuyper, H., Batenburg, T. van, & Hemker, B. (2009). *Technisch rapport eerste meting van COOL$^{5-18}$ in het derde leerjaar van het voortgezet onderwijs*. Groningen/Arnhem: GION/CitoCOOL