UMEÅ UNIVERSITET

Dimensions of validity

Studies of the Swedish national tests in mathematics

Anna Lind Pantzare

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Hörsal E, Humanisthuset,
Fredagen den 30 november, kl. 10:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Gordon Stobart,
Department of Curriculum, Pedagogy and Assessment, University College London, London, England.

Department of Applied Educational Science
Educational Measurement

Organization
Umeå University
Department of Applied
Educational Science

Document type
Doctoral thesis

Date of publication
9 November 2018

Author
Anna Lind Pantzare

Title
Dimensions of validity – Studies of the Swedish national tests in mathematics

Abstract
The main purpose for the Swedish national tests was from the beginning to provide exemplary assessments in a subject and support teachers when interpreting the syllabus. Today, their main purpose is to provide an important basis for teachers when grading their students. Although the results from tests do not entirely decides a student's grade, they are to be taken into special account in the grading process. Given the increasing importance and raise of the stakes, quality issues in terms of validity and reliability is attracting greater attention. The main purpose of this thesis is to examine evidence demonstrating the validity for the Swedish national tests in upper secondary school mathematics and thereby identify potential threats to validity that may affect the interpretations of the test results and lead to invalid conclusions. The validation is made in relation to the purpose that the national tests should support fair and equal assessment and grading. More specifically, the focus was to investigate how differences connected to digital tools, different scorers and the standard setting process affect the results, and also investigate if subscores can be used when interpreting the results. A model visualized as a chain containing links associated with various aspects of validity, ranging from administration and scoring to interpretation and decision-making, is used as a framework for the validation.

The thesis consists of four empirical studies presented in the form of papers and an introduction with summaries of the papers. Different parts of the validation chain are examined in the studies. The focus of the first study is the administration and impact of using advanced calculators when answering test items. These calculators are able to solve equations algebraically and therefore reduce the risk of a student making mistakes. Since the use of such calculators is allowed but not required and since they are quite expensive, there is an obvious threat to validity since the national tests are supposed to be fair and equal for all test takers. The results show that the advanced calculators were not used to a great extent and it was mainly those students who were high-achieving in mathematics that benefited the most. Therefore the conclusion was that the calculators did not affect the results.

The second study was an inter-rater reliability study. In Sweden, teachers are responsible for scoring their own students' national tests, without any training, monitoring or moderation. Therefore it was interesting to investigate the reliability of the scoring since there is a potential risk of bias against one's own students. The analyses showed that the agreement between different raters, analyzed with percent-agreement and kappa, is rather high but some items have lower agreement. In general, items with several correct answers or items where different solution strategies are available are more difficult to score reliably.

The cut scores set by a judgmental Angoff standard setting, the method used to define the cut scores for the national tests in mathematics, was in study three compared with a statistical linking procedure using an anchor test design in order to investigate if the cut scores for two test forms were equally demanding. The results indicate that there were no large differences between the test forms. However, one of the test taker groups was rather small which restricts the power of the analysis. The national tests do not include any anchor items and the study highlights the challenges of introducing equating, that is comparing the difficulty of different test forms, on a regular basis.

In study four, the focus was on subscores and whether there was any value in reporting them in addition to the total score. The syllabus in mathematics has been competence-based since 2011 and the items in the national tests are categorized in relation to these competencies. The test grades are only connected to the total score via the cut scores but the result for each student is consolidated in a result profile based on those competencies. The subscore analysis shows that none of the subscores have added value and the tests would have to be up to four times longer in order to achieve any significant value.

In conclusion, the studies indicate that several of the potential threats do not appear to be significant and the evidence suggests that the interpretations made and decisions taken have the potential to be valid. However, there is a need for further studies. In particular, there is a need to develop a procedure for equating that can be implemented on a regular basis.

Keywords
national tests; validity; interrater reliability; standard setting; subscores; test development

Language
English

ISBN
978-91-7601-936-8

ISSN
1652-9650

Number of pages
61 + 4 papers