Studies in Educational Evaluation 48 (2016) 10–18

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Constructing validity arguments for test combinations

Saskia Wools [a,*], Theo J.H.M. Eggen [a,b], Anton A. Béguin [a,c]

[a] Cito, tav Saskia Wools, Amsterdamseweg 13, 6814 CM Arnhem, The Netherlands
[b] University of Twente, The Netherlands
[c] AQA, UK

ARTICLE INFO

Article history:
Received 23 April 2014
Received in revised form 6 November 2015
Accepted 7 November 2015
Available online 24 November 2015

Keywords:
Validity
Validation
Argument-based approach
Competence assessment programs
Assessment

ABSTRACT

The argument-based approach to validation has been widely adopted in validation theory. However, this approach aims to validate the intended interpretation and use of a single test or assessment. This article proposes an extension of the argument-based approach for validation of multiple tests. This extension is illustrated with the validation of a competency assessment program (CAP). This CAP was validated in collaboration with a quality manager of an educational program. In this case study, it became apparent that this approach fosters an in-depth evaluation of the assessment program and that the approach appears suitable for validation efforts of competency assessment programs. The approach guides validation research from a more general perspective, but also guides more detailed validation efforts.

Validity is often regarded as one of the most important aspects of tests, and although the concept is still under debate (Lissitz, 2009), it is commonly agreed that a test or test score should be valid and reliable (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Especially when high-stakes decisions are made on the basis of test scores, it is necessary to conduct an extensive investigation of the validity of tests or test scores. In education, high-stakes decisions, such as diploma decisions, are rarely based on a single test result. More often, several tests of a different nature are combined into one high-stakes decision (Baartman, Bastiaens, Kirschner, & vander Vleuten, 2007). In this situation, we might not be interested in the validity of a single test or test score, but we would like to be convinced of the validity of the decision. One practical example where test results are combined into one decision is when a competency assessment program (CAP) is used (Baartman, Bastiaens, Kirschner, & vander Vleuten, 2006). Very often, these CAPs are designed to evaluate several aspects of professional competence. The results of the individual components of the CAP are combined to decide whether a student is minimally competent to serve as a starting professional; the student then receives a diploma on the basis of that decision. Therefore, in addition to the validation of the single elements, the decision as a whole needs to be valid. Another example of a single decision informed by a combination of tests is the measurement of growth. Such a program aims to measure one construct, on multiple occasions to identify progress. In this article, an assessment program is defined as a combination of multiple tests or test scores combined into one decision, this could be to measure a multifaceted construct, but can also aim to measure one construct in different ways or on different occasions.

The argument-based approach to validation (Kane, 2006, 2013) has been widely adopted (Brennan, 2013; Lissitz, 2009; Moss, 2013; Newton, 2013; Sireci, 2013) in discussions on validity and validation theory. However, this approach aims to validate the intended interpretation and use of a single test or assessment. In educational practice, tests and assessments are often combined into one decision. To validate this decision, it is possible to validate all parts individually. If these assessment elements are individually considered, we might conclude that some are not sufficiently valid when used in isolation. For example, when only a part of a construct is included in an assessment. However, when these individual assessments are combined with other tests, they might result in a valid decision about students. Therefore, when validating combined tests and assessments, our validation theory must support this. More specifically, the approach to validation should aim to gather validity evidence of the combination as well as evidence of the validity of the individual parts. Therefore, the purpose of this article is to propose an extension of the argument-based approach to validation to guide the validation efforts for decisions based on multiple tests. This extension is illustrated with the validation of an actual CAP. In the next section, the extension is presented after a description of the original approach to validation. This is followed by a description of some examples in which this

* Corresponding author. Tel.: +31 26 3521039.
E-mail address: saskia.wools@cito.nl (S. Wools).

http://dx.doi.org/10.1016/j.stueduc.2015.11.001
0191-491X/

extension can be applied. One of these examples is described in greater detail through an extensive case study. The article then concludes with some remarks on the use of the extended approach, some limitations of this approach, and suggestions for further research.

Theoretical framework

Validity is concerned with the appropriateness of interpretations and uses of test scores (Sireci, 2009), and validation studies are conducted to determine this. These studies aim to gather evidence of a specific interpretation and use of test scores rather than studying the appropriateness of test scores in a broader sense. Kane's (2006, 2013) argument-based approach to validation delineates the intended interpretation and use as one of the main activities to identify assumptions and inferences that are crucial for this intended interpretation and use. Because when the intended interpretation and use are specified, the underlying inferences that seem to be questionable guide us towards the kind of validity evidence that is most needed. As Kane (2013, p. 9) puts it:

> Under the argument-based approach, it is not the case that "almost any information gathered in the process of developing or using a test is relevant to its validity" (Anastasi, 1986, p. 3) or that validation is "a lengthy, even endless process" (Cronbach, 1989, p. 151). The evidence needed for validation is that needed to evaluate the inferences and assumptions in the IUA [*interpretive and use argument*].

The argument-based approach to validation described by Kane elucidates a general framework for validation efforts. Until now, this approach to validation is described for certification testing (Kane, 2004), language testing (Llosa, 2008; Chapelle, Enright, & Jamieson, 2010) and competence assessments (Wools, Eggen, & Sanders, 2010). These tests are all single tests that result in single scores. However, many assessments are used in combination with other tests or measures, especially in educational contexts. This paper therefore aims to extend the argument-based approach for the validation of one assessment to a framework for the validation of multiple tests. Furthermore, it aims to provide an example of validation by means of the argument-based approach for an assessment program that results in a high-stakes decision.

The extension of the argument-based approach is meant for the validation of assessment programs, for example, test combinations for certification purposes whereby complex professional competencies are assessed or test combinations used to assess growth and monitoring of learning progress. The proposed framework is useful for all assessment programs where several test scores or observations are aggregated into one decision. But when multiple decisions are made, the validity of each decision should be determined individually.

The argument-based approach to validation

In this section, the argument-based approach to validation proposed by Kane (2006, 2013) is summarized. Further, the proposed extension of the approach for the validation of assessment programs is described.

The argument-based approach distinguishes two phases: a development stage and an appraisal stage. In the development stage, the intended interpretation and use of test scores are explicitly stated by constructing a, so called, interpretive argument. This argument is shaped as a train of thought that helps with making inferences that more explicitly underlie the assessment. The inferences are categorized using the same model. The actual components of the model are selected on the basis of the intended interpretation and use of the validated assessment.

The basic form of the model, as described by Kane consists of five inferences. The terminology used in this original description could be associated with large-scale standardized tests. Because of the context of this article within educational assessment and competence assessment programs, in some cases other terms are introduced. When different terms are used, the original wording is added in italics. The first inference distinguished in the argument-based approach, or scoring inference, relates to the observed performance of a candidate in a performance test. An evaluation of this observed performance leads to an observed score. Within the generalization inference, this observed score can be generalized to an expected score over the test domain (*universe score*). This test domain represents the universe of tasks that includes all possible tasks. The tasks within the test domain are derived from a competence domain (*level of skill*). This competence domain consists of a written description of the competence or ability of interest. In the interpretive argument, the expected score over the test domain is extrapolated to the competence domain and subsequently to the practice domain within two extrapolation inferences. The practice domain represents the domain about which we would like to make a decision (*target domain*) and is in accordance with the intended interpretation and use of the test. Based on the expected score over the practice domain, a decision can be made in the decision inference.

In short, these inferences are identified (Wools et al., 2010) as follows:

1. Evaluation of the observed performance yielding an observed score.
2. Generalization of the observed score to the expected score over the Test Domain.
3. Extrapolation from the Test Domain to the Competence Domain.
4. Extrapolation from the Competence Domain to the Practice Domain.
5. Decision about readiness for practice.

Every inference included in the interpretive argument must be justified. Therefore, Kane (2006) suggests that within an inference the underlying assumptions are made explicit as part of the interpretive argument. Once the inferences and assumptions are specified, validity evidence to support or reject them should be gathered. Evidence can be both empirical and analytical. Empirical evidence is gathered through trial administrations of the test and (statistical) analyses on the collected data. Analytical evidence is constructed during the development of the test and includes, for example, reports on the rationale of item construction (Wools et al., 2010).

After evidence has been collected and structured according to the interpretive argument, the second stage commences. In this appraisal stage, the evidence is evaluated within a validity argument. In contrast with the interpretive argument, a validity argument is not structured according to a prescribed model. It aims to give an integral evaluation of the appropriateness of the evidence (Kane, 2006) and is shaped in a way that fits this purpose. In this stage, the most questionable assumptions and inferences and the claims that can be easily checked (Cronbach, 1988) are first evaluated; however, assumptions and inferences that are most relevant in relation to the intended interpretation and use are also prioritized. Furthermore, relevant alternative interpretations or rebuttals on the current claims can also indicate sources of evidence needed. Kane (2013, p. 11) describes that "most of the inferences within an interpretive argument are presumptive in a sense that they can establish a presumption in favor of the conclusion but do not establish it definitively". When someone chooses to challenge the presumption this results in a shift of