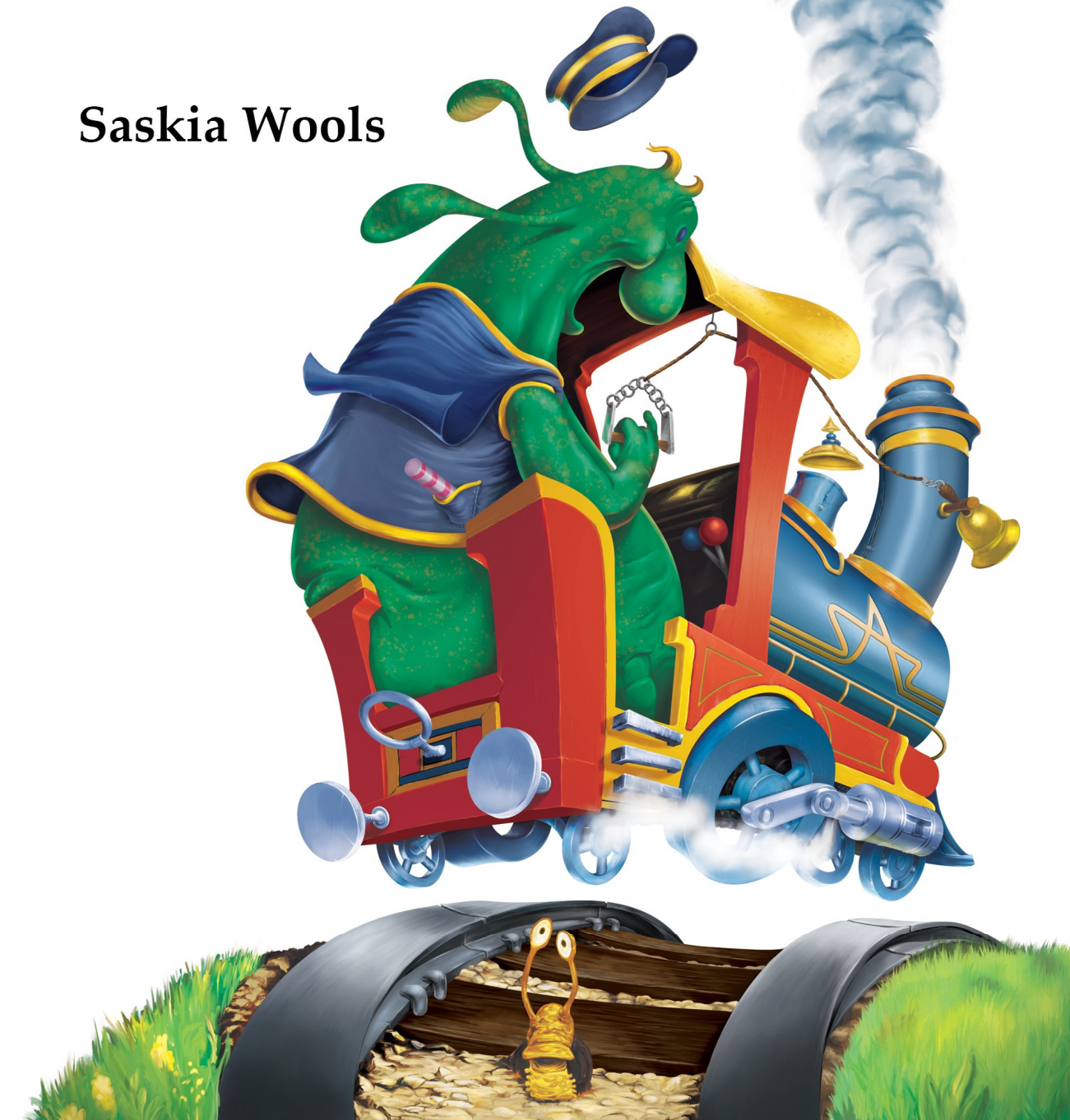
All About Validity

An evaluation system for the quality of educational assessment

Saskia Wools



All About Validity

An evaluation system for the quality of educational assessment

Saskia Wools

Graduation committee:

Chairman	prof. dr. Th. A. J. Toonen
Promotor	prof. dr. ir. T. J. H. M. Eggen
Members	dr. L. Baartman
	prof. dr. J. Cohen-Schotanus
	prof. dr. T. Plomp
	prof. dr. K. Sijtsma
	prof. dr. B. P. Veldkamp

ISBN: 978-94-6259-709-9
Printed by Ipskamp Drukkers, Enschede
Cover designed by Henk van den Heuvel, henk@hillz.nl

This research was supported by Cito, Institute for Educational Measurement.

ALL ABOUT VALIDITY
AN EVALUATION SYSTEM FOR THE QUALITY OF EDUCATIONAL ASSESSMENT

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 26 juni 2015 om 12:45 uur

door

Saskia Wools
geboren op 27 juni 1983
te Apeldoorn

Deze dissertatie is goedgekeurd door de promotor:
prof. dr. ir. T.J.H.M. Eggen

Contents

Abbreviations

ABA	Argument-based approach
AEA-Europe	Association of Educational Assessment - Europe
AERA	American Educational Research Association
APA	American Psychological Association
ATP	Association of Test Publishers
CAP	Competence assessment program
COTAN	Dutch Committee on Tests and Testing
DPA	Driver Performance Assessment
EARLI	European Association for Research on Learning and Instruction
EFPA	European Federation of Psychologists' Associations
ETS	Educational Testing Service
GPA	Grade Point Average
havo	Senior general secondary education
hbo	Higher professional education
ITC	International Testing Committee
IUA	Interpretation and Use Argument
JCTP	Joint Committee on Testing Practices
mbo	Secondary vocational education
NCME	National Council on Measurement in Education
PISA	Programme for International Student Assessment
po	Primary Education
QEA	Quality Evaluation Application
RCEC	Research Center for Examination and Certification
vmbo bb	Pre-vocational education – basic track
vmbo-gt/tl	Pre-vocational education – combined/theoretical track
vmbo-kb	Pre-vocational education – advanced track
vwo	Pre-university education
wo	University education

Chapter 1

General introduction

At all levels of education, tests and assessments are used to gather information about students' skills and competences. This information can be used for decisions about groups of students or about individual students. When individual students are of interest, the decisions made on the basis of test results are of importance during these students' educational careers (Schmeiser & Welch, 2006), for example, when assessment results are used to inform teachers about students' progress on a particular learning goal. Based on this information, a teacher could decide to provide a student with additional learning material to ensure that every concept is grasped. Another example of test use is when results are used by an admissions council to decide which students should be accepted to fill limited college program places. Test results can also be used to evaluate whether students achieved the learning objectives of a study program and whether they should be awarded a diploma.

These different uses of test results require different assessment instruments. Therefore, when tests or assessments are constructed, design choices should be made dependent on the intended use of the test results. When done properly, all these choices are in coherence with the intended use and will benefit the quality of the decisions that test users would like to make.

It is therefore fundamental to evaluate whether test developers succeeded in their efforts to construct assessments that help users make the right decisions about students (e.g., AERA, APA, & NCME, 1999). In this dissertation, it is argued that evaluations of assessment quality should consider the intended use of assessments. When an assessment is used, for example, to certify students who are ready to serve as medical professionals, it should comply with different quality criteria from when it is used to classify students into groups that receive different amounts of instruction. The reasons are two-fold: first, because the stakes of both assessments are very different. Therefore, we need to be more certain of our decision in the first example (certification for practice) than in the second example (different instruction). Second, the actual purpose of the assessment is different, and we might want to evaluate whether the assessment serves its intended purpose. In the first example, we would like to know that students who pass the test are those who are most likely to be successful at performing their job. In the second example, we could evaluate whether the differentiated instruction will lead to better learning outcomes for all students.

In educational measurement, quality evaluation is often done by means of evaluation systems that include guidelines, standards, or quality criteria (Wools, Eggen, & Sanders, 2010). These evaluation systems, are however, not

flexible in use dependent on the purpose or intended use of the assessment that is being evaluated. Often, evaluation systems use the same criteria for all assessments, independent of their purpose. Noteworthy, in some systems, the norms for these criteria differ in relation to the stakes of the test, but the criteria remain the same (Wools, 2012).
This dissertation engages in a description of a design-based research project whose aim is to develop an evaluation system for the quality of tests that evaluates educational assessments dependent on its intended use. This means that not only do the norms of this evaluation differ according to the purpose of a test, the actual criteria with which an assessment needs to comply also differ. To do so, an argument-based approach to quality is introduced. In the literature, this approach is described in the context of validity and validation (Kane, 2013).

Validity is one of the most important quality aspects of assessments (AERA, et al., 1999). It is often defined as the extent to which a test score is appropriate for the intended interpretation and use of the test (e.g., Kane, 2013). To evaluate the validity of test scores, one should gather validity evidence to show the appropriateness of the interpretation and use – a process also known as validation. According to this definition, validity is at the most plausible and is not to be seen as a dichotomous property of tests. In other words, validity is to be interpreted as a continuous property of test score interpretation as opposed to a test being a valid or invalid measuring instrument. The definition stated here could be seen as a consensus definition (Newton, 2012). The actual definition and scope of validity are under constant debate. Newton and Shaw (2014, pp. 176–178) summarize this debate by identifying at least four broad camps: liberals, moderates, traditionalists, and conservatives. *Liberals* extend validity to the overall evaluation of testing policy (e.g., Moss, 2007; Kane, 2013). *Moderates* consider validity to be an evaluation of technical adequacy of testing policy (AERA, et al., 1999). Messick (1998) and Shepard (1997), both *traditionalists*, conclude that test score meaning and test score use are inseparable, thus restricting the definition of validity to the technical evaluation of measurement-based decision-making procedures. Finally, the *conservative* camp believes that validity should only involve the technical quality of measurement procedures. These researchers (e.g., Borsboom & Mellenbergh, 2007; Cizek, 2012; Lissitz & Samuelsen, 2007) argue that validity only concerns test scores and that decision-making should not be of interest to validation research.

In quality evaluation, the liberal view on the concept of validity is less controversial. This view gives the intended interpretation and use of test scores a central position in the discussion and is concerned with the overall evaluation of testing policy. This is also reflected in the recently updated version of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) in which most of the chapters reference the intended use and interpretation of test scores as being the guiding principle of test quality. This means that the choice of the quality criteria should be based on the intended use and interpretation of test scores. This rather flexible view of quality leaves us with a challenge when it comes to (external) evaluation or audits. When quality can be interpreted in different ways, what then does an auditor need to evaluate?
One way to solve this challenge is to ensure that the intended use of an assessment is made explicit. In this way, all those involved with the evaluation of the assessment have the same intended use in mind. Furthermore, when evidence is presented to demonstrate the suitability of an assessment for a particular purpose, it is weighted against the intended use of the assessment, which is stated in advance. This particular approach is also used in validation studies and is rigorously described by Kane (2006, 2013) as the argument-based approach to validation. The original argument-based approach includes two steps: (1) specify the intended interpretation and use of test scores and (2) present evidence that supports or rejects the suitability of test scores for this interpretation and use. When it comes to quality evaluation, a third step is added (Wools et al., 2010): (3) evaluate the presented evidence and decide whether the assessment is fit for purpose.

As mentioned earlier, this dissertation provides a description of a design-based research project that intends to develop an evaluation system, which includes an argument-based approach to quality. Design-based research projects are meant to structure and guide product development, which is, in turn, founded on a theoretical framework (McKenney & Reeves, 2012). By definition, these projects are iterative in nature. In the following chapters, emphasis is placed on building the theoretical framework that serves as a basis for the design phase. In the design phase, design principles were derived from the theoretical framework, and they define the scope of the product being developed. The design principles formulated in this project were translated into a prototype, which was subsequently evaluated against the theoretical framework and the design principles. In the following stages, the prototype was adjusted and evaluated several times until the product was ready. Usually, in every stage, the

goals and hypothesis for evaluation are set, and the methods for evaluation are chosen accordingly. This exemplifies a typical difference in this research method as opposed to other scientific research approaches: in a design-based research project, it is highly recommended that one changes the scope or direction of the project during the study, not afterwards.

Outline

In this project, we borrowed a theoretical framework from the validity and validation literature to be extended to quality evaluation. Three chapters of this dissertation purport to describe the argument-based approach to validation from different angles to exemplify its usefulness in quality evaluation. Chapter 2 starts with describing the argument-based approach to validation and adds a stage, which focuses on quality evaluation, to this approach. In the second part of the chapter, the extended approach is demonstrated in a driver performance assessment for adults.

In Chapter 3, the argument-based approach is exemplified in a very common situation occurring in an educational context – combining multiple assessments into one decision – for example, when multiple assessments are combined into one diploma decision or when several assessments are combined to show growth in ability level. Chapter 3 starts with a theoretical description of the argument-based approach to validation in the context of assessment programs. The theoretical description is then exemplified by validating an assessment program in a Dutch social worker college program.

Following the two extensions of the argument-based approach to validation, Chapter 4 aims to put the approach to use in a complex situation. This chapter purports to show the advantages of the argument-based approach in gaining understanding about the quality of assessments. Furthermore, it shows that the argument-based approach facilitates researchers and policymakers in deciding whether particular design choices contribute to the quality of a decision made within an assessment program. To do so, the chapter focuses on a new national assessment program in arithmetic in the Netherlands. It was identified that the most important claims relate to the comparability of the individual components of the assessment program. Therefore, data was used to evaluate the comparability and to verify the claims made within the program.

In Chapters 2, 3, and 4, the argument-based approach to validation is described from different angles. Chapter 5 focuses on the extent to which the argument-based approach is adopted by researchers when validating tests and

assessments. To do so, a systematic literature review is performed that identifies sources of evidence presented by researchers when reporting on validation efforts. This study reports on the amount of validity evidence presented in journals. Furthermore, it shows that the sources of validity evidence presented differ, to some extent, on the basis of the intended use of the test scores. This latter finding is in accordance with the philosophy of the argument-based approach to validation and its extension to quality.

Chapter 6 starts off with a comparison of currently available evaluation systems for the quality of assessments. This comparison shows a large variety in evaluation systems and their scope. It also implies that it is not relevant to add another evaluation system to this list, rather, it seems useful to provide a system that can include other systems. Therefore, the formulated design principles point towards software that supports quality evaluation from a procedural point of view, that includes an argument-based approach to quality, and that incorporates other evaluation systems.

As a final step of this design-based research project, a prototype of the software was developed and evaluated. The Quality Evaluation Application (QEA) is described in Chapter 7. This online application can be used to build quality arguments according to the argument-based approach to quality. Furthermore, a section of the software is dedicated to quality evaluation. Chapter 7 provides an in-depth description of the system and consists of a description of two evaluation studies performed during the development of the software. These evaluation studies consisted of focus groups that responded to the software. The first group evaluated the first version of the software, which was adjusted on the basis of the results of this evaluation. The adjusted version was then evaluated by a second focus group. The current version of the software is ready to be evaluated in a broader context where test publishers and auditors can both use the software for their own evaluation practices.

This dissertation ends with a discussion on overarching topics that relate to this study but that were not yet addressed in earlier chapters. This includes, for example, a reflection on the usability of design-based research approaches in educational research and comments on the heated discussion on the definition of the concept of validity.

About this dissertation

This dissertation has been written over the course of several years. As time changed, so did a change of terminology in educational sciences. In 2006, Kane described his argument-based approach to validation with an interpretive argument and a validity argument. Newer insights on his part resulted in him changing his terminology in 2013 to an interpretive and use argument (IUA) and a validity argument. The original interpretive argument and the IUA are the same, except that the name changed to ensure that everyone was clear that this interpretive argument also included the intended use of test scores. In this dissertation, both terms are used interchangeably, and the decision was made against changing the terminology since several chapters had already been published or had been submitted for publication with the 'old' terminology.

Another pair of interchangeably used terms is test and assessment. The cultural difference in the appropriateness of the word assessment in the educational context, as opposed to tests being related to psychology, is not widespread. Therefore, the choice was made to use both words interchangeably in order to facilitate readability. Noteworthy, however, in both instances, it is meant to be related to the evaluation situations in an educational context unless otherwise specified.

As a final point, it should be noted that most chapters in this dissertation have been published or have been submitted for publication and are therefore readable on their own. This inevitably results in some overlap and redundancy in the dissertation. However, it was always the intention to keep the description of the theoretical framework in line with the perspective of the chapters. This means that depending on the purpose of a chapter, different elements of the theoretical framework are emphasized, or sometimes, elements are left out completely when they were deemed unnecessary for understanding the chapter.

References

American Educational Research Association (AERA). American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington: American Psychological Association.

American Educational Research Association (AERA). American Psychological Association (APA), National Council on Measurement in Education (NCME).

(2014). *Standards for educational and psychological testing* (2014 ed.). Washington: American Psychological Association.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton and M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–115). New York: Cambridge University Press.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods*, *17*(1), 31–43.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

McKenney, S., & Reeves, T. (2012). *Conducting educational design research: What, why and how*. London: Routledge.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research*, *45*(1-3), 35–44.

Moss P. A. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470–476.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), 1–29.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment.* London: Sage.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Washington DC: American Council on Education.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practices, 16*(2), 5–8.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63–82.

Wools, S. (2012). Towards a comprehensive evaluation system for the quality of tests and assessments. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 95–106). Enschede: RCEC.

Chapter 2

Evaluation of validity and validation by means of the argument-based approach

Abstract:

Validity is the most important quality aspect of tests and assessments, but it is not clear how validity can be evaluated. This article presents a procedure for the evaluation of validity and validation which is an extension of the argument-based approach to validation. The evaluation consists of three criteria to evaluate the interpretive argument, the validity evidence provided, and the validity argument. This procedure is illustrated with an existing assessment: the driver performance assessment. The article concludes with recommendations for the application of the procedure.

Keywords: competence assessment, validity, validation, argument-based approach, evaluation

Chapter previously published as:
Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8*, 63–82.

Theoretical Framework

Introduction

One of the current trends in education is the shift towards more competence-based education (Baartman, Bastiaens, Kirschner & Van der Vleuten, 2007). In the Netherlands, for example, the ministry of education decided that all vocational education institutes must formulate their curriculum according to principles of competence-based education which has led to concomitant changes in learning outcomes. Whereas students used to be taught knowledge and skills separately, they are now acquiring competences in which knowledge, skills, and attitudes are integrated. Also in an international context attention for competencies is increased, in the international programme of student assessment (PISA) for example, cross-curricular competencies are assessed (OECD, 2004).

One of the implications of this change in educational emphasis is an increased use of competence assessments such as performance assessments, situational judgement tests, and portfolio assessments (Baartman, Bastiaens, Kirschner & Van der Vleuten, 2006). These new modes of assessment have been introduced to monitor and assess competence acquisition. Since decisions made on the basis of assessment results can often have serious consequences for individuals, the quality of the assessment instruments needs to be determined to ensure that the right decisions are made.

The evaluation of the quality of assessments is currently at the centre of attention (Anderson Koenig, 2006). Guidelines, standards, and review systems are available to evaluate the quality of assessments or tests. Guidelines are the least prescriptive and only offer guidance in the evaluation process. Standards, such as the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), are more prescriptive but rely on self-regulation for compliance to them (Koretz, 2006). Review systems, lastly, are used to conduct an external evaluation of quality and consist of indicators and criteria to decide whether the quality of an assessment is sufficient. Measurement experts and assessment developers are seeking ways to enforce compliance with guidelines, standards, and review systems (Elliott, Chudowski, Plake, & McDonnel, 2006). Review systems are particularly useful as the results of an evaluation presented in terms of adequate or inadequate makes it possible to attach consequences to these results.

One major condition that needs to be met before compliance to any standard can be enforced is the availability of a widely accepted evaluation system. This evaluation system is not available yet. Despite several attempts to develop or revise standards or other existing evaluation systems, there are still several issues that remain unresolved. These relate to the quality aspects of educational assessments in general, as well as competence assessment more specifically.

One of these issues is the validity of assessments. While validity is the most important quality criterion for any form of assessment, it has thus far been operationalized mainly around the use of standardized tests. Nevertheless, validity is just as important for competence assessments (Messick, 1994). Despite the importance of validity, criteria that can be used for the evaluation of validity of competence assessments are not yet available. Therefore, new criteria to evaluate validity and validation of competence assessments need to be developed.

Validity is basically about the interpretations assigned to test scores rather than the scores themselves (Kane, 1992). Interpreting a test score involves explaining the meaning of a score and making the implications of the scores explicit. The process of evaluating the appropriateness of these interpretations is called validation. In the present article, validation is distinguished from validity: the term validity refers to the use of test scores, whereas validation refers to an activity. As Borsboom, Mellenbergh, and Van Heerden (2004) state: 'validation is the kind of activity researchers undertake to find out whether a test has the property of validity'.

During the evaluation of validity the validation process will also be evaluated, because of the importance of the validation process for establishing validity. However, to ensure a sound evaluation of validity and the validation process, it is preferable that the process is standardised in some way. Therefore a standardised procedure in which the evaluation of validity and validation are integrated is recommended in order to enhance the possibilities of a structured evaluation. The argument-based approach developed by Kane (1992; 2004; 2006) describes a framework that enhances standardisation of the validation process. In the present study, this approach is extended with an additional 'evaluation phase' which consists of newly developed criteria for the evaluation of validity and the validation process.

The procedure for the evaluation of validity and validation based on the argument-based approach is illustrated using a competence-based driver assessment. The driver performance assessment is not administered in an educational setting but has great resemblance with performance assessment in

vocational education. And since the development and validation of this particular assessment are aligned with principles of the argument based approach, this driver assessment is suitable for the illustration provided. In this article, the argument-based approach to validation will be presented first, followed by the criteria for the evaluation of validity and validation. The competence-based driver assessment used in the application of the argument-based approach will then be described. The evaluation of the driver assessments' validation will be used to demonstrate the proposed procedure for the evaluation of validity and validation. The article concludes with recommendations derived from the illustration of the procedure.

Argument-based approach to validation

The argument-based approach consists of two phases: the development stage in which an assessment is developed and an appraisal stage in which the claims being made in the development stage are critically evaluated. During the development stage, inferences and assumptions inherent to the proposed interpretation of assessment results are specified within an interpretive argument. This interpretive argument can be seen as a chain of inferences that are made to translate a performance on a task into a decision on someone's abilities or competences. Figure 2.1 displays an example of inferences that can be included in an interpretive argument.

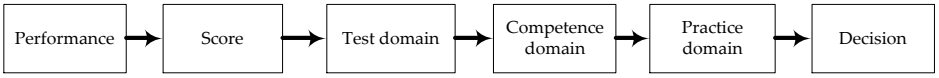


Figure 2.1: Example of inferences in an interpretive argument

This chain of inferences makes the proposed interpretation of an assessment score more explicit by clarifying the steps that can be taken to extrapolate examinees' performances on an assessment to a decision on their level of competence. The first inference relates to a performance on a task that is translated into a numerical score. This observed score is then generalised to a test domain score which represents all possible tasks that could have been presented to examinees. The test domain score is subsequently extrapolated to a score on a competence domain, which entails an operationalization of the competence that is being measured. Within the next inference, the score is extrapolated towards a practice domain. In competence assessments, the practice domain will often be a real-life situation that candidates can be confronted with in their future professional life (Gulikers, 2006). Building on

this final extrapolation, the last inference can lead to a decision on the examinees' level of competence.
When the assessment is fully developed and the interpretive argument is specified, a critical evaluation of the claims being made within the interpretive argument should be made. This critical evaluation takes place in the appraisal stage during which the assumptions stated in the development stage are validated with both analytical and empirical evidence. The analytical evidence could entail, for example, conceptual analyses and judgements on relationships between the test domain, competence domain, and practice domain. Most of the analytical evidence has already been generated during the development stage. The empirical evidence consists, for example, of evidence on the reliability of an assessment. This kind of evidence is gathered in validation studies that are designed to answer specific research questions which are derived from the need for specific empirical evidence. The results of these studies and the analytical evidence are combined and integrated into a validity argument.

Toulmin

Each inference can be seen as a practical argument in which the claim that is made in the preceding inference serves as the starting-point for the next inference. Figure 2.2 represents the form of the arguments and presents a datum, claim, warrant, backing and rebuttal (Toulmin, 1958; 2003). This model is used later in this article to present the inferences within the interpretive argument for the driver performance assessment.

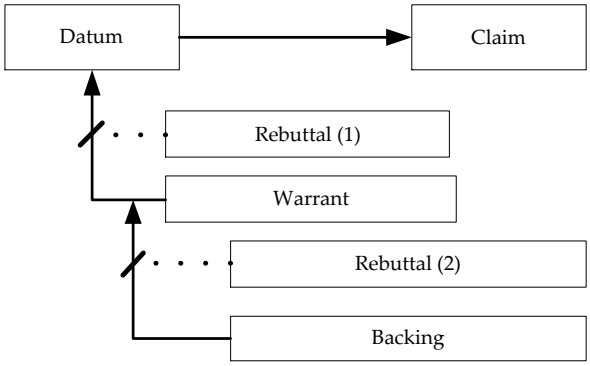


Figure 2.2: Toulmin's model for arguments.

The basis of an argument is the distinction between the *claim* we want to establish and the facts, which is *data*, that serve as the foundation of the claim. Once the data is provided, it may not be necessary to provide more facts that

can serve the claim. Moreover, it is important to state how the data leads to the claim that is being made. The question to be asked should not be 'what have you got to go on?', but 'how do you get there?'. Providing more data of the same kind as the initial data is not appropriate to answer this latter question. Therefore, propositions of a different kind should be raised: rules or principles. By means of these rules or principles, it can be shown that the step from original data to the claim is a legitimate one. The rules and principles will thus function as a bridge from data to claim. These bridges are referred to as *warrants* and are represented in Figure 2.2 by an arrow. Because the warrants possess neither authority nor currency, the distinction between data, on the one hand, and warrants, on the other, is not an absolute distinction since some warrants can be questioned. Supporting warrants are other assurances that are referred to as *backing*. Lastly, Toulmin mentions a *rebuttal*, which indicates circumstances in which the general authority of the warrant would have to be set aside. A rebuttal provides conditions of exception for the argument and is represented in Figure 2.2 by a dotted line and a forward slash.

Criteria to evaluate the validity and validation of assessments

To evaluate the validity and validation of assessments, a third stage is added to the development and appraisal stages in the argument-based approach: the evaluation stage. Within this stage, criteria are applied to evaluate the interpretive argument and the validity argument. The criteria for evaluation of the validity and validation of assessments are based on theories on the evaluation of informal and practical arguments as it is not possible to evaluate practical arguments in the same way as formal arguments. Because the available evidence is often incomplete and sometimes questionable, the argument as a whole is, at best, convincing or plausible.

The proposed evaluation takes place on two levels: first, two (conditional) criteria should be met to ensure a sound validation process, and then a third criterion to ensure validity is applied. The aim of the first criterion is to evaluate the quality of the interpretive argument because it is preferable that the inferences chosen correspond with the proposed interpretation of the assessment. Furthermore, it is necessary that the interpretive argument and its inferences are specified in detail because, in that case, gaps and inconsistencies are harder to ignore. Therefore, it is desirable that each inference includes at least one backing, one warrant and one rebuttal. These aspects are covered in the first criterion:

1. Does the interpretive argument address the correct inferences and assumptions?

The second criterion takes the validity evidence presented into account by evaluating each inference as proposed in theories on the evaluation of informal logic (Verheij, 2005). When arguments are evaluated in formal logic, it has to be decided whether an argument is valid or invalid. However, in the evaluation of Toulmin arguments, an 'evaluation status' is introduced. To determine the evaluation status of the individual inferences, the first step is to evaluate the assumptions and statements included in the argument individually and decide whether each statement or assumption is accepted, rejected, or not investigated. The second step is to assign an evaluation status (Verheij, 2005) to the inference as a whole: justified, defeated, or unevaluated. This decision is made based on decision rules that underlie Toulmin's arguments:

- The evaluation status is justified when the warrant(s) and backing(s) are accepted and the rebuttal(s) are rejected.
- The evaluation status is defeated when a warrant of backing is rejected or when a rebuttal is accepted.
- The evaluation status is unevaluated when some statements are not investigated and it is still possible for the inference to become justified.

The theory described here can be used to decide on the second criterion:

2. Are the inferences justified?

The third criterion concerns an evaluation of the outcomes of the validation process. Owing to the condition that the first two criteria must be met before the third criterion is applied, it is already determined that the right inferences were chosen and it is also established that the inferences are justified. The next step is to evaluate whether the validity argument as a whole is plausible. For this we need to take all evidence into account to decide whether the argument is strong enough to convince us of the validity of the assessment. The third criterion that will be answered is:

3. Is the validity argument as a whole plausible?

Evaluation of validity and validation

The procedure for the evaluation of validity and validation is illustrated for the driver performance assessment (DPA). First, the main elements of this driver

assessment are presented, followed by the interpretive and validity argument. In the discussions of the second part, more detailed information about the driver assessment is provided when needed for an understanding of the arguments.

Driver Performance Assessment (DPA)

The driver performance assessment (DPA) is an on-road assessment reflecting a competence-based view on driving. This assessment instrument can be used to establish drivers' driving proficiency and is appropriate for learner-drivers as well as experienced drivers and is meant to guide further driver training. The DPA is used as part of an on-road training session. Part of this session consists of driving without intervention from the driving instructor who observes the driver's driving skills. The driver is instructed to drive along a representative route through five different areas: residential access roads inside and outside built-up areas, roads connecting towns inside and outside built-up areas, and highways. In order to judge the drivers' proficiency, a matrix was developed in which the tasks, areas, and criteria of the DPA were combined. Table 2.1 presents these elements schematically.

As shown in Table 2.1, the DPA distinguishes various driving tasks that are categorized under five main tasks: preparing for driving, making progress, crossing intersections, moving laterally, and carrying out special manoeuvres. Each task can be performed in each area. And all these driving tasks are judged against five performance criteria: safe driving, consideration for other road users, facilitating traffic flow, environmentally responsible driving, and controlled driving. The driving instructor is expected to score each cell of the matrix on a rating scale from 1 (very unsatisfactory) to 4 (optimal).

The driving instructors who acted as assessors were trained to carry out the performance assessments. During three 3-hour workshops, they learned how to use the scoring rubric and tried to reach consensus on the interpretation of the performance criteria. Furthermore, the instructors assessed 12 video-clips showing critical parts of the task performance of four different drivers.

Table 2.1: Schematic presentation of elements

		Criteria:				
Area:	**Task:**	**Safe driving**	**Consideration for other road users**	**Facilitating driving flow**	**Environmentally responsible driving**	**Controlled driving**
- Residential access road (inside built-up area) - Residential access road (outside built-up area) - Roads connecting towns (inside built-up area) - Roads connecting towns (outside built-up area) - Highways	Preparing for driving	1 - 4	1 - 4	1 - 4	1 - 4	1 - 4
	Making progress	1 - 4	1 - 4	1 - 4	1 - 4	1 - 4
	Crossing intersections	1 - 4	1 - 4	1 - 4	1 - 4	1 - 4
	Moving laterally	1 - 4	1 - 4	1 - 4	1 - 4	1 - 4
	Special manoeuvres	1 - 4	1 - 4	1 - 4	1 - 4	1 - 4

Developing interpretive and validity arguments

The DPA validation studies were carried out to gather validity evidence. However, an interpretive or validity argument has never been developed. To formulate these arguments, the DPA was studied thoroughly and an interpretive argument was developed. Subsequently, the validation studies were examined and all validity evidence was classified within the inferences of the interpretive argument. All evidence selected was then summarised into a final validity argument.

Illustration

This section contains the illustration of the procedure for the evaluation of validity and validation for the DPA. First, the interpretive argument for the DPA is addressed, then the validity argument is presented, and finally the

application of the criteria for the evaluation of validity and validation is described.

Interpretive Argument for the DPA

The proposed interpretation of assessment scores is specified within the interpretive argument. This specification consists of a description of the inferences that are made to extend the score on a test performance to draw conclusions about a candidate's proficiency.

With the DPA, a decision on a driver's driving proficiency in a real-life situation is made. Real-life driving is described in terms of 'driving competence' which is operationalized into possible driving tasks. The drivers, however, only perform a selection of these tasks. The performance on this selection of tasks is expressed by a DPA score. The reasoning mentioned here, is formalised into the same inferences of the interpretive argument presented in the description of the argument-based approach that includes a scoring inference, a generalization inference, two extrapolation inferences, and a decision inference. Figures 2.3 through 2.7 present the five inferences for the DPA structured according to the Toulmin Model presented earlier. The first inference will be presented in detail; the following inferences will be summarised.

When the DPA is administered, the driver drives along a route that is indicated by an instructor. The instructor assesses the performance and, with the use of score rubrics and scoring rules, allocates a numerical score to the driver's performance. This procedure is framed within an argument according to the Toulmin model (Figure 2.3). Figure 2.3 shows that it is possible to allocate a numerical score to a performance on the DPA, which is the transition from datum (performance) to claim (score). The score is allocated by qualified raters (warrant) but this will only lead to a consistent score when these raters reach a sufficient level of agreement (rebuttal 1). Furthermore, a rater can only allocate a score when score rubrics and scoring rules are available (backing). It is clear that these scoring rubrics and rules can only be of help when they are applied accurately (rebuttal 2).

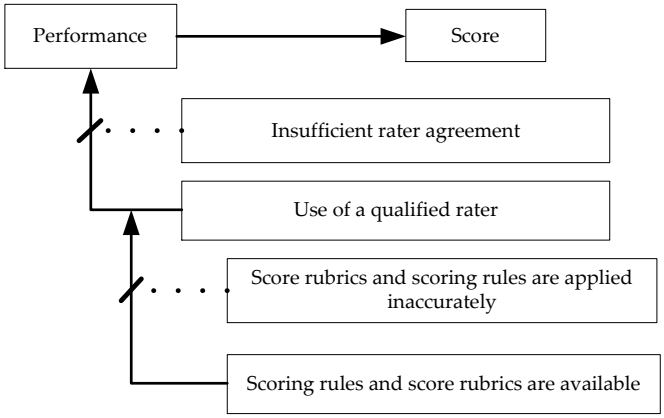


Figure 2.3: Scoring inference - evaluation of the observed performance on the DPA yielding an observed score

The second inference (Figure 2.4) leads from the observed score on the DPA towards an expected score over the test domain which consists of all possible tasks that could be administered. The observed score can be generalized into a score for the test domain when the administered tasks are representative for the test domain regarding the content. Furthermore, to allow generalization, the sample of tasks needs to be large enough to control sampling error. Note, however, that the claim that the backing supports the warrant is only valid when the conditions in which generalization is evidenced are the same as during a regular administration.

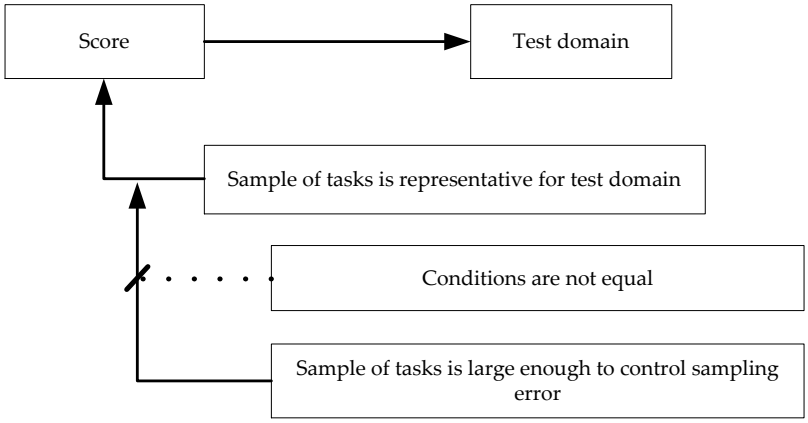


Figure 2.4: Generalization inference - generalization of the observed score on the DPA to the expected score over the test domain

The extrapolation from the test domain to the competence domain of driving is accounted for in inference 3 which is presented in Figure 2.5.

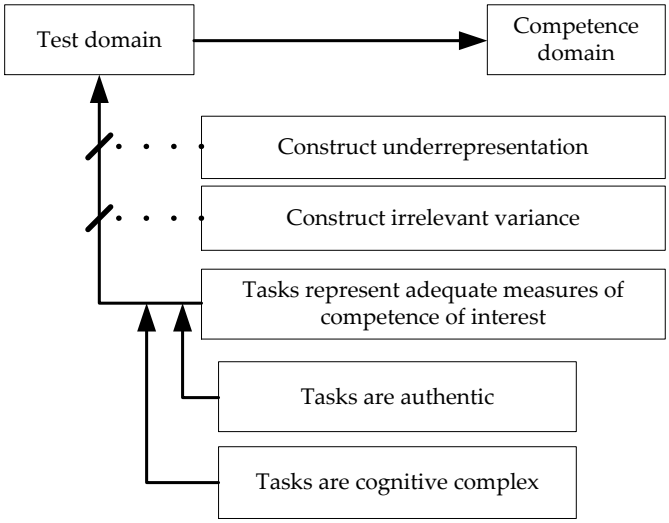


Figure 2.5: First extrapolation inference - extrapolation from the test domain to the competence domain of driving

For this extrapolation, it is necessary that the tasks generate a performance that is a reflection of the competence described. The DPA requires drivers to drive in an on-road situation without the intervention of the instructor. This means that the task performance of the driver provides direct evidence of the driver's driving competence. There are two threats (rebuttals) to extrapolation included in this argument: construct underrepresentation and construct irrelevant variance. The term construct underrepresentation indicates that the tasks that are measured in the DPA fail to include important dimensions or aspects of driving competence. The term construct-irrelevant variance means that the test outcomes may be confounded with nuisance variables that are unrelated to driver competence. Besides these threats, there are also two indicators added that serve as backing for the representation of the competence domain: the tasks should be authentic and the tasks should be cognitively complex. Authentic means that the tasks should be as similar as possible to 'real-life driving'; and cognitively complex means that the tasks should address all cognitive processes that are necessary when driving.

Figure 2.6 presents the fourth inference which is the extrapolation from the competence domain of driving to the practice domain of driving. This inference can be made because the competence domain is based on a theoretical

description of the practice domain of driving (real-life driving). Of course, this is only possible when the competence domain is not too narrow and all relevant aspects of driving and all conditions under which drivers perform are included. Within the operationalization of the practice domain, the 'critical driving situations' should be made explicit. These critical situations relate to crucial aspects of driving that can contribute to distinguishing between different levels of driving proficiency.

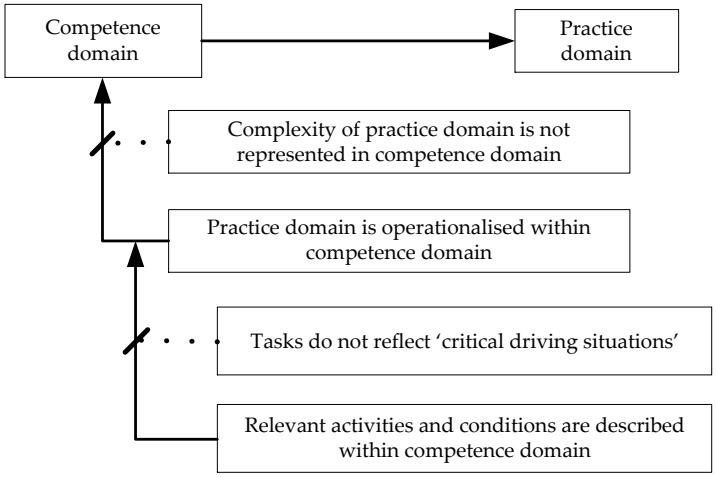


Figure 2.6: Second extrapolation inference - extrapolation from the competence domain of driving to the practice domain of driving

The inference in Figure 2.7 shows that decisions can be made based on the practice domain of driving. A cut-off score is available to make a decision on the driver's driving proficiency. This cut-off score supports the last inference since it is established with a standard-setting procedure in which certain levels of performance in the practice domain are connected to certain DPA scores. The rebuttals that are distinguished in this inference relate to the correctness of the cut-off score and to the appropriateness of the standard-setting procedure that lead to a cut-off.

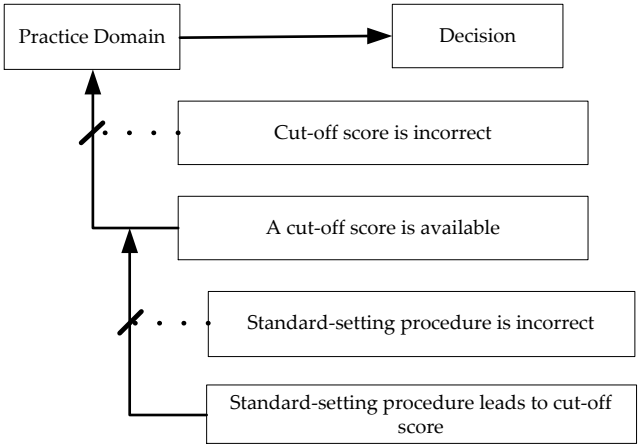


Figure 2.7: Decision inference: from the extrapolation to the practice domain of driving it is possible to make decisions on the driver's driving proficiency

Validity argument

It is argued within the validity argument that administering the DPA leads to valid decisions on drivers' driving proficiency. Evidence to support the validity argument is gathered during the development phase as well as during the appraisal phase of the argument-based approach to validation. A verbal summary of this validity argument is presented below. Note that a validity argument is based on available evidence and is generally written by test-developers to convince test users of the validity of test scores. Whether this is a legitimate claim will be investigated during the evaluation of the validity argument.

For the DPA, the scoring inference - from performance to score - can be made since experienced driving instructors who received additional training are responsible for scoring the performance. To score the performance on a rating scale ranging from 1 to 4, the instructors use score rubrics in which driving tasks are judged against the five criteria mentioned in Table 2.1. There is also a detailed scoring manual available to support instructors during the scoring. Inter-rater agreement coefficients, that is, Gower coefficients (1971), were calculated for every criterion to indicate instructors' mastery of the assessment procedure. Inter-rater agreement coefficients were between .74 and .82, which can be considered at an acceptable level.

The generalization inference - from score to test domain - can be made since the DPA distinguishes different areas and different driving tasks and therefore the content domain is covered. The instructors select the number and diversity of

tasks by choosing a representative route that will take approximately one hour. Test-retest reliability is also estimated to determine whether the sample of tasks is large enough to control bias that occurs through an incorrect sample of tasks. With a correlation of .80 between the first DPA-score and the second DPA-score for the same drivers, test–retest reliability for the DPA is sufficient.

The expected score on the test domain can be extrapolated to an expected score on the competence domain because the tasks within the DPA are related to the description of driving as a competence. Therefore, the first extrapolation inference can be made. The tasks are authentic since the learner-drivers are supposed to perform driving tasks in an on-road situation. The tasks are also cognitively complex since they are divided over different levels of task performance distinguished for driving: the strategic level, the tactical level, and the operational level. These levels of task performance correspond with the description of driving competence found in the literature on this topic. Because the tasks are authentic and cognitively complex, it is possible to extrapolate the expected score on all possible tasks to the competence domain.

The competence domain resulted from a description of the practice domain, therefore the second extrapolation inference - from competence domain to practice domain - can be made as well. The literature on driving practice (Hatakka, Keskinen, Gergersen, Glad, & Hernetkoski, 2002) is used to form a competence-based view of driving. Furthermore, during the development of the DPA, traffic and driving experts were consulted to make sure critical driving situations were accounted for.

The validation studies show that the decision inference which states that the expected score on the competence domain leads to a decision, can be made. The DPA-scores were related to the results of the final driver exam in order to compare the DPA-scores to an external criterion. It appeared that the mean DPA-scores for learner-drivers who passed the final exam are significantly higher than the learner-drivers who failed the final exam. Therefore, it is assumed that learner-drivers who are less competent receive lower DPA scores than learner-drivers who are more competent. To make a distinction between these groups, a cut-off score is set based on the external criterion and with this cut-off score the percentage of misclassifications is calculated. Learner-drivers are misclassified when they receive a DPA score below the cut-score but pass the final exam and the other way around. The percentage of misclassified learner-drivers in the validation study performed was 35.9%. Since the DPA is a formative instrument, this percentage of misclassified drivers is still acceptable.

In conclusion, we argue that, based upon evidence presented within the validity argument, it is possible to make valid decisions on drivers' driving proficiency based on the administration of the DPA.

Evaluation

After the interpretive and validity arguments are specified, the three criteria for the evaluation of validity and validation can be applied. The first criterion evaluates the interpretive argument and the specified inferences, the second criterion evaluates the evidence presented, and the final criterion evaluates the validity argument.

Criterion 1: Interpretive argument

The number of inferences included in the interpretive argument reflects the complexity of the DPA. Since the purpose of the assessment is to decide on a learner-driver's driving proficiency, it is necessary to extrapolate a performance to a practice domain. Therefore, at least four inferences must be made: scoring inference, generalization, extrapolation, decision. The extrapolation inference that consisted of two parts, firstly, from test domain to competence domain and, secondly, from competence domain into practice domain, does not affect the completeness of the interpretive argument negatively.
Another aspect of this criterion is the amount of detail in which the inferences are specified, because, as mentioned before, it is harder to ignore gaps and inconsistencies within an interpretive argument when it is specified in detail. Table 2.2 shows whether a backing, warrant or rebuttal is present for every inference.

Table 2.2: Number of backings, warrants, and rebuttals included in inferences

Inference	Backing	Warrant	Rebuttals
Scoring inference	1	1	2
Generalization inference	1	1	1
Extrapolation inference (1)	1	2	2
Extrapolation inference (2)	1	1	2
Decision inference	1	1	2

Since the number of inferences included in the interpretive argument for the DPA is sufficient, and every inference includes at least a backing, warrant, and rebuttal, this criterion is satisfied.

Criterion 2: Evaluation of evidence

The second criterion for evaluation is applied to evaluate whether the evidence presented is plausible and whether the inferences are coherent. Therefore, the evaluation status for each inference is determined as described earlier. The first inference of the interpretive argument of the DPA, from performance to score, is justified as is shown in Figure 2.8.

First of all, the warrant (W) is accepted since the raters are certified driving instructors with many years of experience. The backing (B) is accepted as well because of the availability of detailed scoring rubrics. Both rebuttals are, however, declined. The first rebuttal (R1) is declined because the rater agreement reached an acceptable level, that is, a mean of Gower coefficients above .70. The second rebuttal (R2) is declined since a great deal of effort has been put into a correct application of the scoring rules and rubrics during the training of the raters and because there is no evidence that the raters applied the scoring rules and rubrics inappropriately during the scoring of the performance assessment.

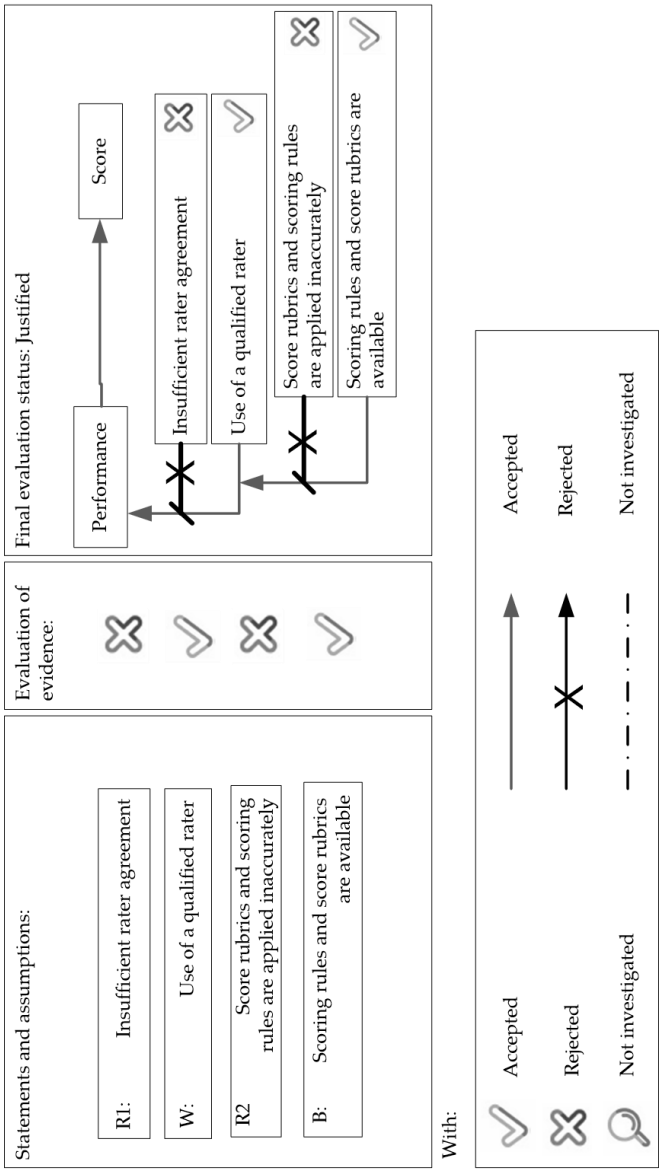


Figure 2.8: First inference of the interpretive argument of the DPA: Justified

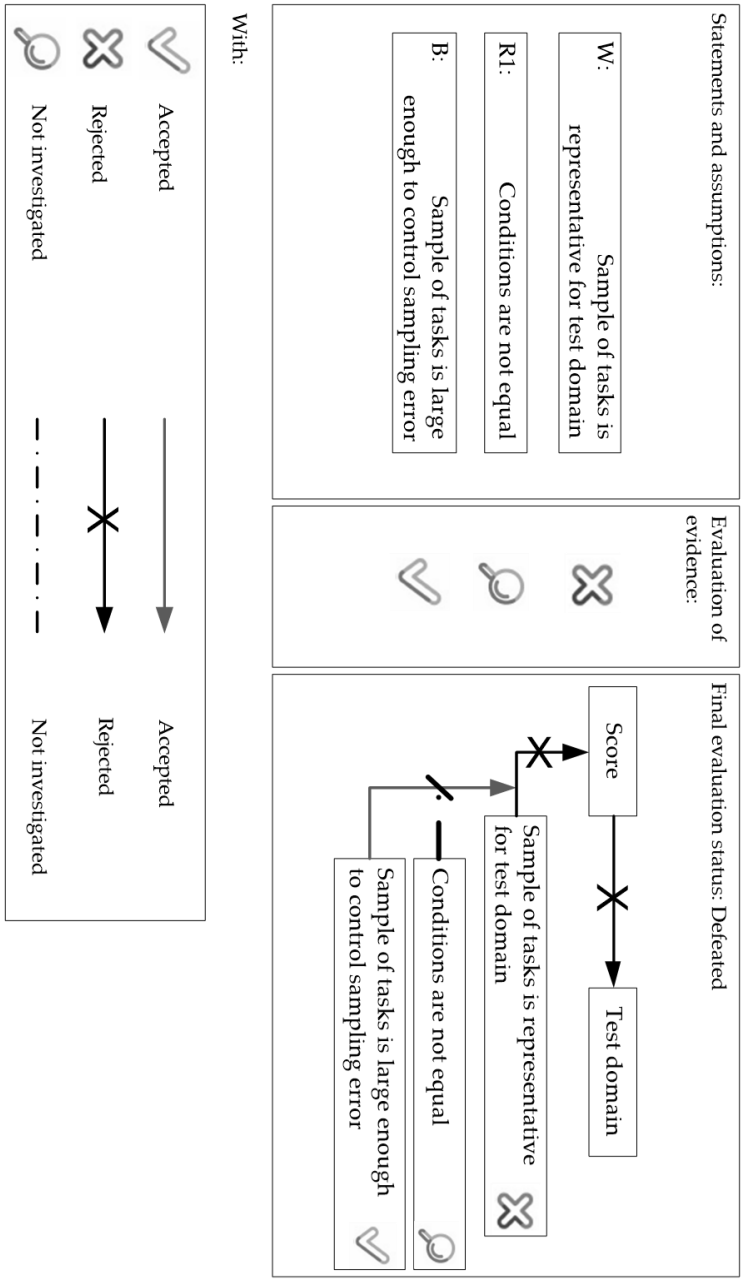


Figure 2.9: Second inference of the interpretive argument of the DPA: Defeated

The generalization inference of the DPA is shown in Figure 2.9. For this inference, the warrant (W) is rejected and therefore the inference has already been defeated despite the fact that the backing (B) is accepted and the rebuttal (R) was not investigated. For the warrant to be accepted, it must be plausible that every candidate performs on at least every task distinguished and within every area distinguished. Since there is no evidence that indicates that this is true, the warrant is rejected. The backing is accepted because of a sufficient test–retest reliability, but because of the rejection of the warrant, this does not change the evaluation status of this inference.

The evaluation status unevaluated is assigned to the first extrapolation inference which is presented in Figure 2.10. Both warrant (W) and backings (B1; B2) are accepted, but both rebuttals (R1; R2) are not investigated. The warrant and backings are accepted based on developmental evidence which means that the tasks were developed by traffic experts and that driving instructors were involved in the development of the DPA. For the inference to be justified, it is necessary that both rebuttals be rejected. When only one rebuttal is accepted, the inference will be defeated.

The second extrapolation inference of the interpretive argument of the DPA, from competence domain to practice domain, presented in Figure 2.6, is justified because both rebuttals are rejected. At the same time, the warrant and backing were accepted because of evidence such as the contribution of traffic and driving experts in the description of the practice domain.

The decision inference, presented in Figure 2.7, remains unevaluated because there is still little evidence for both the backing and the rebuttal on the backing. However, the warrant is accepted since a cut-off score is available. The rebuttal on the warrant is rejected based on the significant differences in mean DPA scores for learner-drivers who passed and failed the final exam.

In conclusion, by applying the second criterion, it appeared that only two inferences were justified. Additional validation research should aim for the validation of the two inferences that are unevaluated. Furthermore, it is necessary to adjust elements of the assessment to justify the inferences that are defeated for the moment.

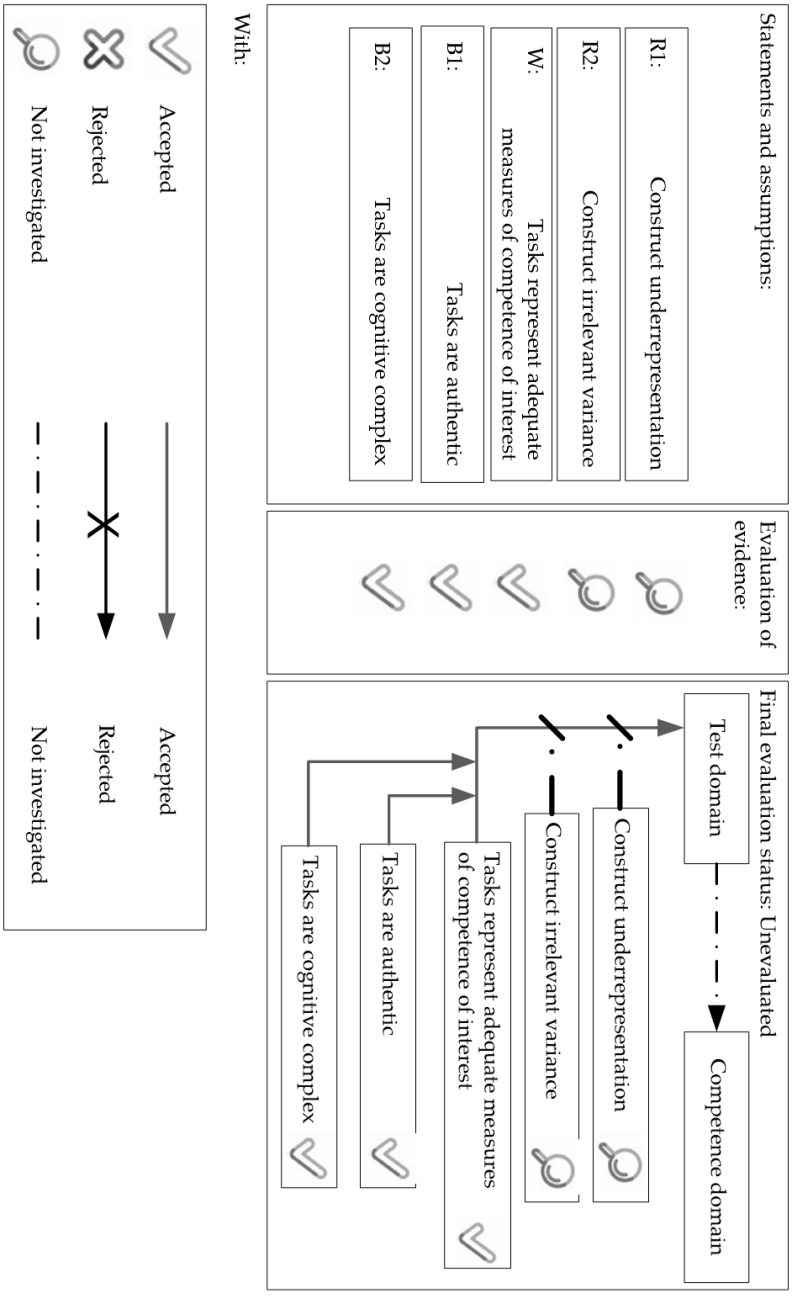


Figure 2.10: Third inference of the interpretive argument of the DPA: Unevaluated

Criterion 3: Evaluation of the validity argument

The interpretive argument and the evidence presented were evaluated through the application of two conditional criteria. The second criterion, however, as described in the previous section, has not been met and, therefore, the third criterion would normally not need to be applied. However, for illustrative purposes, an example of an answer for the third criterion is nonetheless presented. The third criterion focuses on the evaluation of the validity argument: *Is the validity argument as a whole plausible?*

The validity argument as a whole is not plausible because of a lack of evidence for several inferences. The evidence gathered during the development phase is convincing and provides plausible arguments for the validity of the DPA. However, the evidence gathered within the appraisal phase is not convincing. This is not because of the size of the study group (N=91; N=61), but because the validation studies focused particularly on establishing a cut-off score. Studies that focus on estimating reliability or try to establish whether there is construct-irrelevant variance might strengthen the evidence presented.

In addition, it does not add to the plausibility of the validity argument that the goal of the DPA is set to be formative, or, in other words, to present certified and uncertified drivers with evaluative information about their driving proficiency. However, all evidence presented is gathered within groups of learner-drivers. And furthermore, a cut-off score is set to distinguish between candidates that are likely to pass the final driver examination and candidates that are likely to fail. This cut-off score is not consistent with the stated goal of this instrument. It seems that the validity argument actually supports the claim that the DPA is suitable to decide whether a candidate is ready to participate in the final examination instead of providing insight into a driver's strengths and weaknesses to guide further training.

Conclusion

The purpose of this article was to illustrate a new procedure for the evaluation of validity and validation. For this procedure, the argument-based approach to validation was extended with an evaluation phase. Within this phase, criteria are applied to evaluate the quality of the validation process as well as the validity of test results. In this last section some recommendations regarding the application of the proposed procedure for evaluating validity and validation are given. The recommendations relate to the application of the argument-

based approach and more specifically the evaluation phase. The article concludes with suggestions for future research and development.

Argument-based approach to validation

The first recommendation regarding the application of the argument-based approach relates to the construction of an interpretive argument. During the development of the interpretive argument for the DPA, it appeared that it is very complicated to formulate a complete interpretive argument that includes all relevant aspects. Therefore, it is recommended that an interpretive argument should be developed by a development team. This team should include a content expert and a measurement expert to ensure that measurement considerations as well as issues regarding the content of the assessment are accounted for in the interpretive argument.

The second recommendation addresses the availability of analytical evidence. To enhance the strength of a validity argument, it is necessary to account for the analytical evidence during the development phase. It is thus important that every step in the development phase is thoroughly documented.

The third recommendation regarding the application of the argument-based approach concerns the guiding role it can play in validation research. When an interpretive argument is developed with regard to the preceding recommendations, it becomes evident what additional validation studies should aim for during the appraisal stage. That way, it is relatively easy to focus solely on the statements and assumptions that need to be affirmed.

Evaluation phase

During the evaluation phase several elements are evaluated by applying the criteria. It turned out that it is important to distinguish the different phases and criteria. It is, for example, important to evaluate the interpretive argument by means of the first criterion (are the correct assumptions and inferences addressed?) without taking the evidence presented into account. The latter is only evaluated with the second criterion on the justification of the inferences itself. This process should be supported with software designed to guide the validation process, the evaluation process, and to help in presenting the results of both processes.

Furthermore it became apparent that it is necessary to decide what the minimal requirements are for valid assessments. Especially during the evaluation of evidence, it is necessary to define what is good enough. In the illustration that was presented, it remained quite arbitrary when evidence was accepted or

rejected. Therefore, it is recommended that some kind of standard-setting procedure be performed to define the minimal requirements for evidence before conclusions on the quality of validity and validation can be drawn.
The last recommendation relates to the evaluation of the plausibility of the validity argument, the third criterion. It should be discussed whether it is acceptable and desirable for one criterion to be quite judgemental because, despite the fact that the first and second criteria provide explicit decision rules, the last criterion still requires a judgement call.

Where to go from here?

This article addresses a procedure for the evaluation of validity and validation. However, during the application of this procedure, it appeared that this evaluation entails more than just validity and validation. This could have been expected because validity and test quality are highly related (Messick, 1994). Nevertheless, it might be interesting to investigate the possibilities of using the argument-based approach to validation as a framework for the evaluation of tests and assessments in general.
The use of the argument-based approach as a general framework for the evaluation of quality of tests and assessments requires more research. This research should focus on elements within individual inferences. First of all the question whether it is possible for all relevant elements of test quality to be accounted for in the arguments needs to be investigated. Furthermore, before an external evaluation of the quality of the validation process can be performed it is necessary to study, on the inference level, what evidence is essential for assessment experts to accept the claims being made to make sure that the conclusions of the evaluation phase are valid, acceptable, and plausible.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing.* Washington: American Psychological Association.

Anderson Koenig, J. (2006). Introduction and overview: Considering compliance, enforcement, and revisions. *Educational Measurement: Issues and Practice* , *25* (3), 18-21.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, 32, 153-170.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2,* 114–129.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061-1071.

Elliott, S., Chudowsky, N., Plake, B. S., & Mcdonnell, L. (2006). Using the standards to evaluate the redesign of the U.S. naturalization tests: Lessons for the measurement community. *Educational Measurement: Issues and Practice , 25* (3), 22-26.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857-871.

Gulikers, J. T. M. (2006). *Authenticity is in the eye of the beholder: Beliefs and perceptions of authentic assessment and the influence on student learning.* Dissertatie. Heerlen: Open Universiteit.

Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A. & Hernetkoski, K. (2002). *From control of the vehicle to personal self-control; broadening the perspectives to driver education.* Transportation Research Part F, 5, 201-215.

Kane, M. T. (1992) An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement, 2,* 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurment 4th edition.* (pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Koretz, D. (2006). Steps toward more effective implementation of the standards for educational and psychological testing. *Educational Measurement: Issues and Practice*, *25* (3), 46-50.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,*13-22.

OECD. (2004). Learning for tomorrow's world. First results from PISA 2003. Paris: OECD Publications.

Roelofs, E., Vissers, J., van Onna, M., & Nägele, R. (2009). Validity of an on-road driver performance assessment within an initial driver training context. *Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* Montana, 482 – 490.

Roelofs, E., van der Linden, A., Wools, S., Nägele, R., & Vissers, J. (2008). Constructie van een formatief instrument voor de beoordeling van praktische rijvaardigheid [Construction of a formative instrument to assess driving proficiency in practice]. *Paper presented at the Onderwijs Research Dagen 2008*, Eindhoven.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The uses of argument. Updated Edition.* Cambridge: Cambridge University Press.

Verheij, B. (2005). Evaluating arguments based on Toulmin's scheme. *Argumentation*, 19, 347-371.

Chapter 3

Constructing validity arguments for combinations of tests

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.

Keywords: Validity, Validation, Argument-based Approach, Competence Assessment Programs, Assessment

Chapter submitted for publication (December 2014):
Wools, S., Eggen, T., & Béguin, A. (2014). Constructing validity arguments for combinations of tests. *Studies in Educational Evaluation.*

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 (AERA, APA, & NCME, 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, & Van der 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, & Van der 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 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 puts it (2013, p. 9):

> 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. The two phases consist of different activities and emphasis but are not totally distinct. In practice, activities are performed in a For example, when evidence collected in the 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 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, Eggen, & Sanders, 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, Eggen, Sanders, & 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 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. Finally, claims that are easily checked are evaluated (Cronbach, 1988). If the proposed interpretation and use are supported by the evidence and alternative explanations are rejected, the validity argument is concluded by stating whether or not it is valid to interpret test scores in the proposed way (Kane, 2006).
Both stages are described here as being distinct with their own emphasis and purpose, however, in practice both stages can be followed iteratively. For example, when evidence is collected that rejects a major claim made within the interpretive argument, one might decide to change the interpretive argument and the proposed interpretation and use accordingly.

The extended Argument-based approach to validation of assessment programs

When the argument-based approach is used to validate multiple tests, the procedure is the same. Validation efforts continue to be structured in two stages. The content of the arguments does however differ. Since the arguments are shaped according to the claims and inferences being made during test construction, the interpretive argument is shaped differently when multiple tests are combined into one decision. This decision might comprise of a judgment of ability on a multifaceted construct where multiple elements are combined. It can also cover an evaluation of ability on a single construct but with multiple operationalizations to demonstrate this ability. An original interpretive argument reasons from one performance to a decision. The extended interpretive argument can, however, incorporate multiple performances, multiple test domains, and multiple competence domains that are aggregated into one decision. The validity argument also differs since the emphasis of the argument shifts from supporting a single test score interpretation to the aggregation of test scores. Therefore, the issues that become relevant for the validation of an assessment program differ from those that are addressed in a validity argument for a single assessment. To elaborate on these differences, this section addresses both the interpretive and validity arguments.

Interpretive argument of assessment programs

The same elements used within an interpretive argument for a single assessment are also used within an interpretive argument for assessment programs. The types of inferences characterized within the argument-based approach to validation in both cases are as follows: scoring, generalization,

extrapolation, and decision rules (Brennan, 2013). The main difference, however, is that some inferences can be used multiple times to indicate relations within the assessment program. This allows us to construct an argument that fits the rationale of the program, but also to make relations within the program explicit. For example, when a decision is made regarding three competences, we can construct three interpretive arguments by distinguishing three competence domains and connect each of them to one test domain, one score, and one performance. But when a competence domain is operationalized into performance tasks and knowledge items, it enhances the clarity of the argument when two test domains are combined into one competence domain. Furthermore, by explicitly stating that the competence domain is operationalized into two test domains, it guides our validation efforts towards gathering evidence to support or reject this specific design choice.
Figure 3.1 displays an example of an interpretive argument that connects the elements in different ways. In this particular example, three performances are distinguished and lead to one decision. In the remainder of this sub-section, every element is discussed in greater detail. In general, the need to distinguish between one or more elements is dictated by the proposed interpretation and use of the assessment program.

Scoring inference (Performance - Score)

Scoring inferences focus on rules and procedures for obtaining the observed scores that are ultimately used for interpretations and decisions (Brennan, 2013). If multiple tasks are carried out in an assessment program, it should be determined whether a performance should be regarded as "single" or "multiple" instances. A factor that could be considered here, is the amount of time scheduled between different tasks. Because when the tasks are observed at different times, we might regard them separately, for example, when the same ability is measured every year to ascertain students' progress or when different competence-related aspects are tested and measured after a few weeks of training. In contrast, when we are explicitly interested in a student's ability at a specific moment, it might be logical to regard the performance as a single instance. This might be the case even when students need to perform on two parts of a test administered over two days. The second aspect that could influence the choice of performances being perceived as "single" or "multiple" is the type of task that students need to execute. More specifically, when tasks

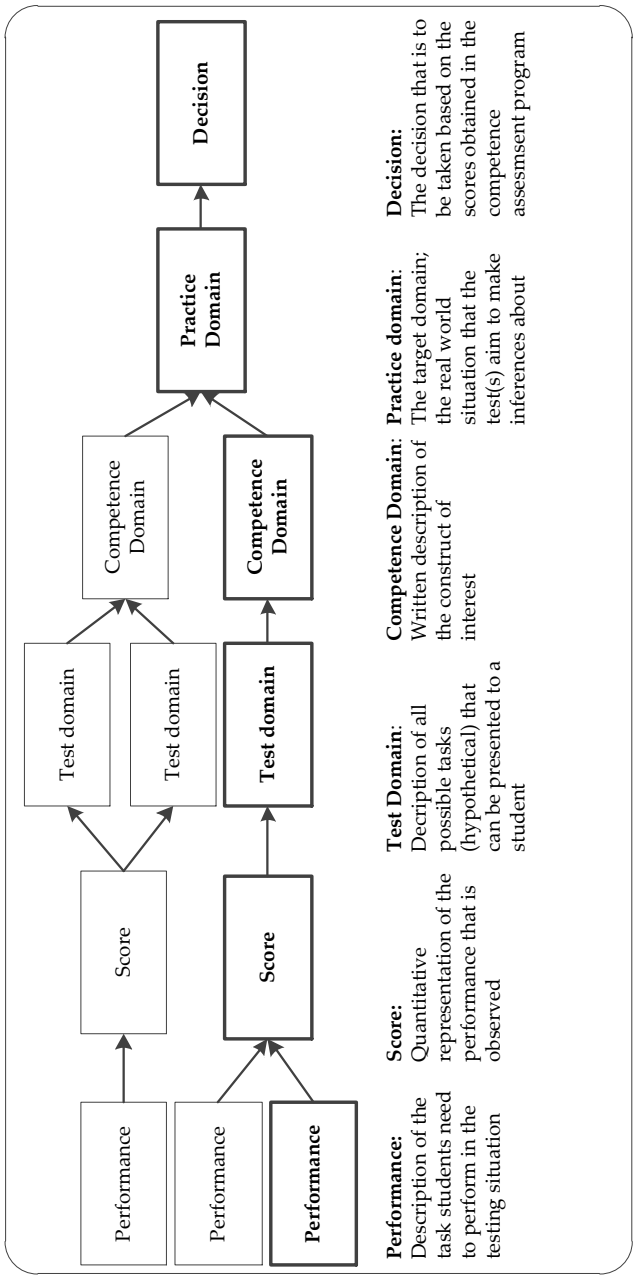


Figure 3.1: Example of an interpretive argument with multiple inferences

are very different in nature, they could be regarded as separate performances even though they are aimed to be generalized into the same competence (Baartman et al., 2007).
An interpretive argument describing an assessment program might combine multiple performances into one score. For example, when scores on two parts of a test are taken together or when a number correct score on a listening test is added to a number correct score on a reading test. The reverse is also possible: two scores can be derived from one performance, for instance, when separate raters rate different aspects of one performance.

Generalization inference (Score - Test Domain)

One might choose to define several test domains that each represent a different operationalization of a construct of interest. For example, in language testing, we could distinguish a test domain that includes all possible reading tasks and one that includes all possible speaking tasks. When these two are distinguished during the construction of the interpretive argument, it then becomes apparent that scores need to be generalized to both test domains and that both contexts are of interest in measuring competence. When these two examples are combined into one test domain, this could indicate that there is a strong relation between the two elements and that, for example, compensation between these elements is possible. However, it could also mean that it is implicitly assumed that every task within the domain is interchangeable. In our language test example, this would mean that we would be able to draw conclusions about a students' language proficiency solely on the basis of speaking tasks, or it could mean that it is assumed that scores obtained by students on the speaking part of the test could be compared to scores obtained by students on the reading tasks.

Extrapolation inference I (Test Domain - Competence Domain)

The hypothetical scores on a test domain are extrapolated into a score on a competence domain. To do so, we claim that test performances can be used to make interpretations about non-test performances in non-test contexts (Haertel, 1999). These non-test performances and contexts can also be described as competence domains, that is, a written description of the construct of interest. These competences are often described in units that focus on a single domain or competence. However, when the domain of interest is specifically divided into separate components, two competence domains can be derived from one construct, for example, when tests for driving licenses are divided into practical driving skills and theoretical knowledge of traffic rules.

Extrapolation inference II (Competence Domain - Practice Domain)

The practice domain is also referred to as target domain. Assessments aim to provide scores that can be interpreted as a measure of competence or ability on this particular domain. In vocational education, this domain often refers to the real world in which professional competences have to be demonstrated (Gulikers, 2006). Since this is the most thorough description of the target domain, and it is related to the real world, there is only one practice domain within an interpretive argument.

Decision inference (Practice Domain - Decision)

Within an interpretive argument, the decision that is of interest depends on the proposed interpretation and use of the assessment program. Or in other words, the intended use dictates the kind of decision that needs to be made. In this approach only one decision can be specified and supported with evidence. When multiple decisions or purposes are distinguished, multiple interpretive arguments need to be constructed.

Validity argument of assessment programs

When a validity argument for assessment programs is constructed, it differs from a validity argument for single tests in such a way that emphasis is put on other elements. The elements that are addressed within validity arguments for assessment programs probably address claims that are specifically related to test-assessment combinations. When the most questionable aspects of an assessment program concern a single aspect of the program, it is more appropriate to first validate this single element in isolation. When this element is proven to add to the body of knowledge supporting the validity of the test scores, the other elements that are related to the combination of tests should also be supported by evidence. To explain the validity argument in greater detail, this section will present questions that could be posed within validity arguments for different assessment programs. Note that the specific questions posed in a validity argument are always prompted by the interpretive argument that fits the intended interpretation and use and the specific assessment situation. The following assessment programs serve as an example of situations in which assessment programs might be used and validity arguments should be made.

Validity argument for measuring progress

In this example, the assessment program purports to measure progress over several years of education. Every year, students take a test that includes items covering learning goals for the current year and items covering learning content from previous years. To compare the performances of students over time, vertical linking procedures are often used (Carlton, 2011; Harris, 2007; Kolen & Brennan, 2004, pp. 372–418). These procedures are IRT-based methods that make scale-based interpretations possible even when different tests are used. However, to use these techniques, many data-related assumptions must be met. The interpretive argument for this assessment program emphasizes the growth of students proficiency over time and claims that students who have learned during the intervals between the test moments obtain higher scores than their peers who have not learned as much.
Possible questions that could be raised in the validity argument for an assessment program that measures progress include

- Is the test content suitable for students in the particular grades, or are some items aimed at younger/older students?
- How is progress defined? Are students who learn to do specific tasks better equally awarded in terms of progress as students who learn to do more tasks (but at the same level of complexity)? Are psychometric models in concurrence with this definition?
- Are all statistical assumptions necessary to use vertical linking techniques met? Are the results of these technique sufficiently robust?

Validity argument for measuring complex competences obtained over multiple years of education

This assessment program includes several tests that aim to ensure that students obtain learning goals. Every course within the educational program ends with a test. By combining the results of all tests, it is ensured that all relevant aspects are accounted for in the decision on passing or failing the entire educational program. The combination-rule used for this decision is that students need to pass every course before they can receive a diploma. The interpretive argument emphasizes the fact that all aspects of the learning targets are included in the diploma decision and that students who receive a diploma are sufficiently competent to move on to the next stage. Note that the case study presented in the last part of this article is a specific example of this assessment program.
The validity argument for this example should address questions, such as:

- Are all tests sufficiently reliable to prevent a large number of false negatives?
- Does the program include sufficient tests that aim for the cognitive complexity represented in the practice domain?
- Is the test combination aligned in a way that ensures that several operationalizations are used? For example, does the test combination include multiple choice, open-ended, and performance tests?
- Is the ability demonstrated by students in tests at the beginning of the educational program stable over time? Are we able to make inferences about these performances years after they have been demonstrated?

Validity argument for measuring several abilities in an admissions procedure

When several sources of information are combined into one admissions decision, this could be regarded as an assessment program. These kinds of decisions could, for example, be based on an assessment program that includes the test results of different test developers, an interview with the student, and a motivational letter. The claim made within this assessment program, and that should be specified within an accompanying interpretive argument, is that students who have the best chances of succeeding in the program of interest are selected for admission.
In such an admissions assessment program, the validity argument should address questions, such as:

- Are all sources of evidence equally important, and how are these weighted?
- When different test scores are available, are all scores comparable and equally valid?
- Are inter-rater differences that occur in making the admission-decision accounted for?

The examples discussed in this section were simplified because in educational programs, practical restrictions are often in place, and assessment programs tend to be more complex. To illustrate the possibilities of this approach, the remainder of this article presents its application in the validation of a real-life assessment program.

Case study

This section introduces a real-life case to exemplify the previously described extension of the argument-based approach to the validation of multiple tests. The example of the competency assessment program (CAP) for a vocational study program for social workers is described, including the method used to obtain information and evidence of this example. To develop both an interpretive and validity argument for the social worker CAP, all available documents were collected. Based on these documents, a first draft of the interpretive argument was constructed. This draft was used in an interview with the test quality manager of the social worker track. During this interview, the assumptions underlying the interpretive argument were tested and possible sources of evidence were discussed. It became clear however, that by explicitly discussing the relations within the assessment program and the underlying claims, understanding of possible sources of evidence grew. Furthermore, the quality manager did not only focus on the existing program and its validity but was also inspired to evaluate design choices for a new program. The readily available evidence focused mainly on describing the procedures regarding test evaluations. Other sources of evidence relating to the content of the assessments were not easily accessible.

Given the purpose of this case study as an extensive example, a decision was made to continue the study without collecting the sources of evidence from individual teachers. As opposed to presenting actual evidence, it was regarded more informative and generalizable to pose questions that exemplified different possible solutions than to present a limited number of actual sources of evidence.

The instrument used to illustrate the validation of a combination of tests is a CAP for social workers. This CAP is constructed as part of a four-year study program at the Dutch University of Applied Sciences. The elements of the CAP are validated and administered on several occasions. Within the CAP, three key competences are distinguished: “designing and executing treatment plans”, “working in a professional environment”, and “becoming a better professional”. All key competences are divided into several components and are tested using different assessment forms, such as performance assessments, interviews, and knowledge tests. After four years, the results of all tests are combined into one decision about students’ ability to perform as a starting professional in the field of social work. Note that as described earlier, this case is a specific example of an assessment program for measuring complex competences obtained over multiple years of education.

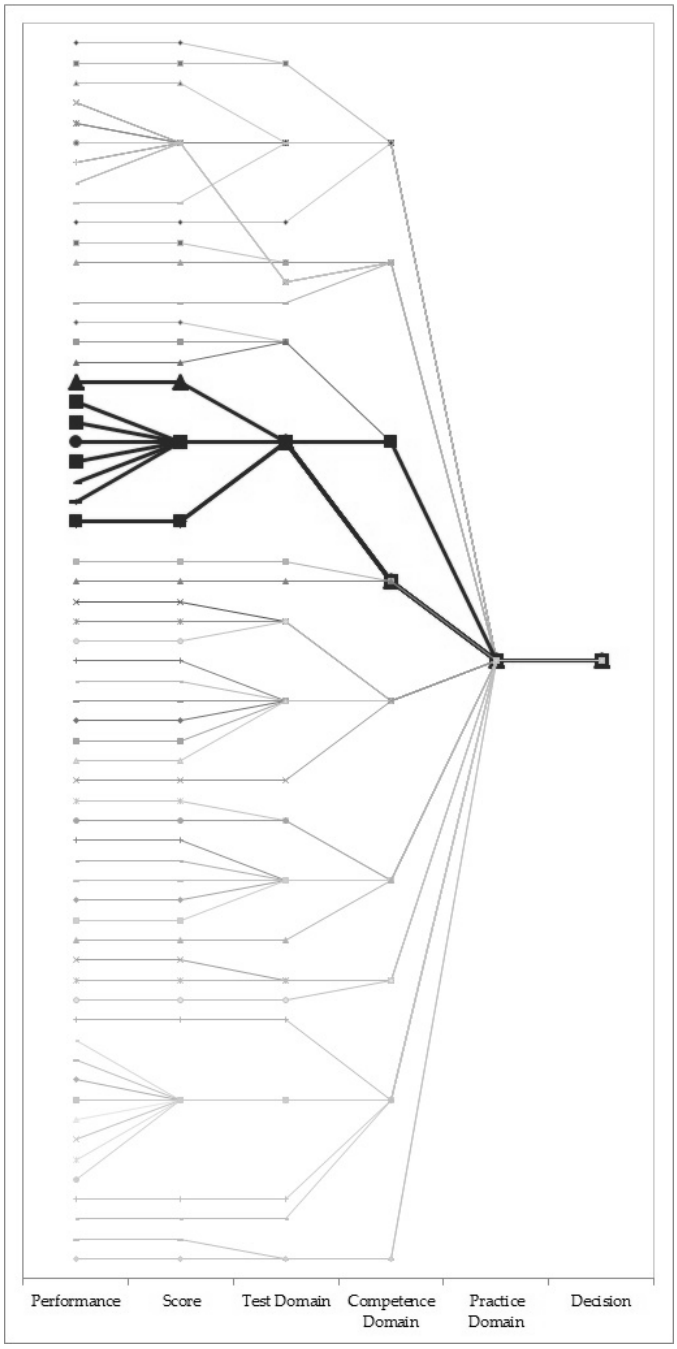


Figure 3.2: graphical representation of the full CAP

Assessment Program for Social Workers

Interpretive argument CAP - social workers

In this CAP (graphically represented in Figure 3.2), students perform in 60 evaluative situations, which are translated into 45 different scores. The performances are very diverse, ranging from writing papers to sitting in a theoretical exam, answering multiple choice questions, and giving an oral presentation. The scores are sometimes given as: sufficient or insufficient. However, scores on a scale from one to ten are also awarded. These scores are then generalized into 23 test domains, and from there, they are extrapolated into nine competence domains. The competence domains are derived from the practice domain and separates the work field of social workers into nine core elements that can be performed in four different working contexts. The test domains, however, describe tasks according to different levels of complexity within these nine core elements. The easiest level is that of orientation where students simply need to know what is going on; the intermediate level is that of reproduction where students need to reproduce the tasks; finally, the expert level is that of production where students can work independently. Every level of task-performance is described within one test domain. The competence domains are all extrapolated into a single practice domain, which leads to a single decision: is a student minimally competent to work as a starting professional in the field of social work?

Inferences (assumptions and evidence)

To show the variety of combinations of assumptions and inferences within the interpretive argument, we will discuss one part of this argument in greater detail. This part consists of inferences that lead to two competence domains and one practice domain. Figure 3.3 shows the structure of the part of the argument discussed here, the inferences that are discussed more elaborate are also bold in Figure 3.2. In this section, we will further describe the inferences in this figure.

Scoring inference (Performance - Score)

In Figure 3.3, we see that students need to perform on eight tasks. The knowledge test (performance A) gives a score of 1-10 points; the reflection paper (performance H) leads to a sufficient or insufficient score. The other performances (performance B-G) are combined into one overall portfolio score.

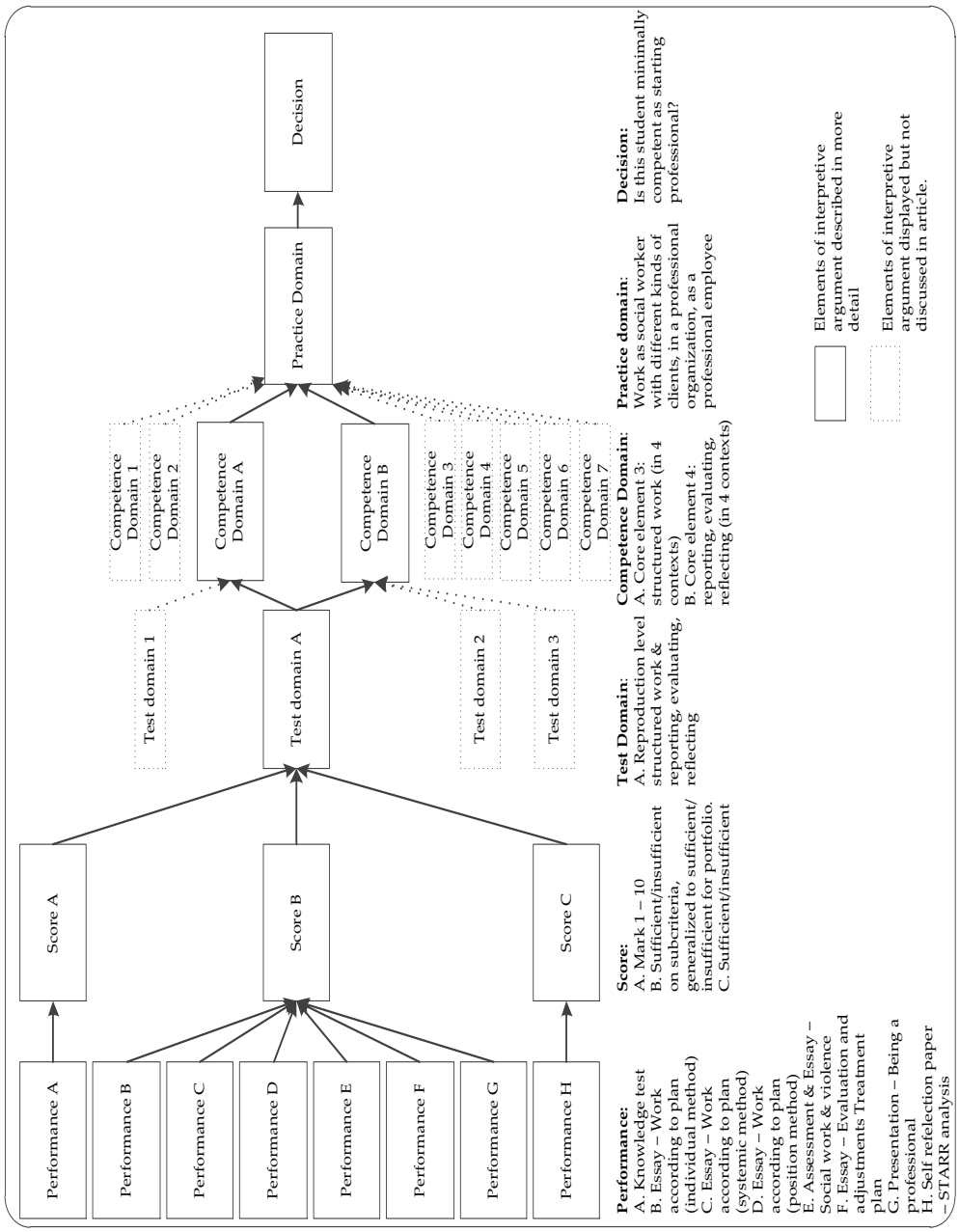


Figure 3.3: Part of the interpretive argument for a 4-year educational program for social workers

This combination is made on the basis of criteria that are scored for every element of the portfolio. Once all elements are scored, the criteria are put together and combined using a scoring rule to report a sufficient or insufficient score on the completed portfolio.

Generalization inference (Score - Test Domain)

The three scores (A, B, and C) are generalized into one test domain. This test domain consist of tasks on the reproduction level that relate to both structured work and reporting, evaluating, and reflecting.

Extrapolation inference I (Test Domain - Competence Domain)

The single test domain is extrapolated into two competence domains. The competences that are taken together in the test domain are described separately as competences (structured work versus reporting, evaluating, and reflecting). The extrapolated scores for the competence domain are not only derived from Test Domain A but are also combined with Test Domains 1, 2, and 3, which contain tasks on different levels of complexity (orientation and production levels). This decision means that each task in the test domain elicits behavior that demonstrates ability in both competences.

Extrapolation inference II (Competence Domain - Practice Domain)

From here, all described competences are combined into one practice domain: the actual work field that a social worker will enter upon finalizing this study program. The observed behavior on the two competences (structured work & reporting, evaluating, and reflecting) are extrapolated to possible behavior in actual work situations.

Decision inference (Practice Domain - Decision)

When a prediction is made on possible behavior in the actual work situation, this should be translated into an answer to the question: is this student minimally competent to advance as a starting professional?

Validity argument CAP - social workers

The validity argument comprises a critical evaluation of the claims and inferences made in the interpretive argument. In this section, the validity argument that fits the previously described interpretive argument for the social worker program is presented (Table 3.1). Since the actual study program did

not have all evidence available and it was not structured according to this approach, this specific validity argument consists of questions that can be posed in relation to the general evidence rather than actual evidence. These questions are derived from assumptions underlying the interpretive argument. Note that the questions serve as an example and that depending on the intended interpretation and use of the assessment program, other questions could be posed or emphasized. Furthermore, it must be stressed that the actual content of a real validity argument depends on the evidence, the goal of the validation efforts, and other case-specific properties. The validity argument presented here is only an example of the structure of a validity argument for CAPs.

The validity argument presented here shows that a significant amount of evidence is needed to evaluate an entire CAP. It also guides us to the areas in which evidence is most needed. Since this particular social worker program uses competences (competence domain) that are designed by an external committee, it might want to focus on gathering evidence on other aspects of the argument. The most prominent inferences that need evidencing are those related to the test domain since only one test domain is constructed from two competence domains. Another important inference for this particular program is the final inference that leads to a decision about students. It should be clear how the performance standard described by the external committee, which consists of content experts, is translated into cut scores, performance indicators, or other tools to guide the decision-making process on students' success or failure.

From the interview on the presented arguments, it became clear that there was especially a lack of attention on the operationalization of the competence domain into a test domain. A significant amount of effort was going into developing items, tasks, and rating schemes, but in in terms of deciding on what kinds of tasks were necessary to elicit the behavior of interest, there was no rationale. By building this arguments, this became very clear since there was no available information for the first extrapolation inference (test domain - competence domain).

Table 3.1: Validity argument for the 4-year educational program for social workers

1A	**Performance A - Score A**
Evidence:	• Are the keys of all multiple choice items correct? • Is the sum score calculated correctly?
1B	**Performance B - G - Score B**
Evidence:	• Is a scoring guide or rating scheme available? • Are raters applying the rating scheme in the same way? • How are different components combined into one score?
1C	**Performance H - Score C**
Evidence	• Does the rating scheme include all relevant aspects of the performance?
1	**Performance – Score**
Summary	Standardized procedures are applied to express three performances into three numerical scores. When several raters are involved, rater agreement must be presented. Furthermore, it needs to be ensured that all raters hold the same interpretation of the rating scheme. It must be made clear that it is necessary to combine several into three scores as opposed to assigning a score for every performance. Lastly, it must be evidence why two performances, in particular (knowledge test & reflection paper), are significantly emphasized and weighted.
2A	**Score A - Test Domain A**
Evidence	• Are the selected items based on a test matrix? • Does the test matrix include all relevant aspects of the test domain? • Is the number of items large enough to control for sampling error?
2B	**Score B - Test Domain A**
Evidence	• Are tasks aimed at the right level of task complexity? • Is the content of tasks related to the diversity of the construct? • Are all relevant aspects of the test domain reflected in tasks that aim for performance at the reproduction level?
2C	**Score C - Test Domain A**
Evidence	• Is it necessary to address one aspect of the test domain in more detail than in others?
2	**Score - Test Domain**
Summary	The test domain consists of two main elements, both of which have to be reflected in the scores. It is necessary to gather evidence on the combination of these two elements into one test domain and on the necessity of this choice. One notable feature is that within the three scores, "reflecting" is emphasized more than the other aspects of the test domain. Although the content of the tasks seems to be on the reproduction level, there are no comparisons made with easier (orientation level) or more difficult (production level) tasks in terms of item difficulty.

3A	**Test Domain A - Competence Domain A**
Evidence	• Are all aspects of the competence domain translated into tasks? • Do the tasks elicit behavior of interest? • Are the tasks developed in a way that content-irrelevant behavior is minimized?
3B	**Test Domain A - Competence Domain B**
Evidence	• Is there a task that isolates the behavior of interest from content-related competence? • Is the content of tasks in concurrence with the ideas of experts on intended behavior?
3	**Test Domain - Competence Domain**
Summary	The tasks are of a different nature: multiple choice items, assignments for papers, and a reflection report. On the basis the competence domain, it is questionable whether students only need to show their competence in writing. However, this might be the case since this test domain only covers the reproduction level.
4A	**Competence Domain A - Practice Domain**
Evidence	• Does the written construct in the competence domain include all relevant aspects of the practice domain? • Is the content of the competence domain accepted by relevant stakeholders, such as future employers? • Is the content aligned with modern views in the profession?
4B	**Competence Domain B - Practice Domain**
Evidence	• Are both content-specific and general competences included in the competence domain? • Is the description of the construct aimed at the right level of difficulty in the practice domain?
4	**Competence Domain - Practice Domain**
Summary	The competence domain is described by a committee comprising relevant stakeholders. This committee decides on the necessary competences and the level of difficulty (in terms of content) that is necessary to be minimally competent as a professional. Documentation regarding this process of involving stakeholders might be evidence that could be presented here.

5	**Practice Domain – Decision**
Evidence	• Are stakeholders involved in setting the performance standards? • Are these standards translated to tasks in a standardized way? • Are those students who pass rightly classified as minimally competent professionals?
Summary	Performance standards are based on the description of the practice domain. It is however not clear how these standards are translated to the individual tasks. Furthermore, evidence should be provided on the future performance of students who passed the program in order to verify the classification accuracy of the program.

Discussion and conclusion

In this paper, we presented an extension of the argument-based approach (Kane, 2006, 2013) to assessment programs with multiple tests and assessments. These assessments are combined into one decision. By combining these assessments into one validity argument, the validity of the decision as a whole can be evaluated. To evaluate the validity of the decision, an interpretive argument is built; it consists of multiple inferences that specify the claims being made within an assessment program. This interpretive argument guides researchers in understanding what element should be emphasized when gathering validity evidence. It is argued that within an interpretive argument for an assessment program, the combination of elements should be most evidenced. Since the specific combination of tests is a key element of an assessment program, this should be expressed in the evidence that is presented on the validity of the decision. Therefore, the validity argument that is constructed consists primarily of evidence that supports or rejects claims that are related to the combination of scores. To exemplify the usability of this theory, three common assessment programs were described, and questions that could be raised within a validity argument were suggested. Furthermore, a case study of an assessment program for social workers was presented as an extensive example of both an interpretive argument and a validity argument.

This case study was constructed in collaboration with a quality manager of the social worker program. Through this collaboration, it became apparent that this approach fosters an in-depth evaluation of the assessment program and that the argument-based approach to validation appears suitable as a guide for validation efforts for competency assessment programs. The approach guides validation research from a more general perspective; it also guides more

detailed validation efforts. The program exemplified here did, for instance, have problems in relation to evidencing the extrapolation of a test domain to a competence domain, which is a more general issue. At a more detailed level, for some tests, it was not clear whether all raters used the rating schemes in the same manner. The main advantage of this approach is that it structures the assessment program in a way that makes conversations on the program possible. It provided stakeholders with the possibility of viewing and discussing the program as a whole and, therefore, to find inconsistencies in the test combinations. In the past, users regarded all tests individually but could not see, for example, that the combination of all tests focused too heavily on writing essays. These conclusions fit the main purpose of the argument-based approach to validation, that is, guiding research to evaluate the validation of tests (Kane, 2013). The usability of this method is not limited to a particular kind of assessment program, to fully use the advantages of this approach it is, however, desirable that all related evidence is easily available. A potential risk of the approach, is that users think of it as a checklist for validity. It is only suitable for identifying gaps in validation research, and that it could not solve particular validity issues. In other words, simply using this approach does not make an assessment program more valid.

A downfall of the approach is the complexity of building the arguments. Because of the complex nature of the approach, it might be necessary to clarify the approach for it to be adopted by practitioners. Regarding simplification, Sireci (2013) suggests an adaptation of the argument-based approach in a way that the underlying theoretical principles are still in place but that the intended use and validity evidence are articulated and presented without the structure of an interpretive argument. To further simplify the approach, Newton (2013) proposes that we lose the distinction between the interpretive argument and the validity argument and only acknowledge the validity argument. In this paper, it is also recognized that the distinction between the two arguments and the prescribed form of the interpretive argument add to the complexity of the theory. However, a distinct interpretive argument also facilitates discussions about the design of the assessment program in relation to the intended interpretation and use of the combined results. At the same time, a distinct validity argument provides us with the opportunity to evaluate the claims made within the interpretive argument from a more distant perspective. Therefore, in this paper, the suggestion is not to simplify the argument-based approach as it is. It is acknowledged though that practitioners need more guidance in the use of the framework for their everyday validation practices.

A possible way to support practitioners in building interpretive and validity arguments is by providing them with tools that can simplify the process. It would be helpful, for example, to develop software to support practitioners in their validation efforts. This software could facilitate the collection and storage of sources of evidence and could help with building an argument over several years and collecting evidence along the way. If such a tool were available for the case study presented here, we would have been able to construct an actual validity argument. For future research, it is necessary to build the interpretive argument during the construction of the assessment program and to collect evidence accordingly. Only in this way can the full possibilities of the extension of the argument-based approach be exemplified.

Until then, it seems that the general practice within schools is ill-equipped for a full implementation of this approach. Therefore, the current use of this theory might be at a stage prior to actual validation, for example, to foster discussions on the design of a competence assessment on a program level and an identification of strengths and weaknesses. When it becomes clear that the quality of a program as a whole can be improved with the use of this approach, it might become more worthwhile to invest time, expertise, and effort in collecting and structuring validity evidence according to this approach.

References

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing.* Washington: American Psychological Association.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, *32*, 153–170.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2,* 114–129.

Brennan, R. L. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores". *Journal of Educational Measurement*, *50*, 74–83. doi: 10.1111/jedm.12001.

Carlton, J. E. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59 – 70). Springer: New York.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference ? *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*. Springer: New York.

Gulikers, J. T. M. (2006). *Authenticity is in the eye of the beholder: Beliefs and perceptions of authentic assessment and the influence on student learning.* Dissertatie. Heerlen: Open Universiteit.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*, 135–70.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating* (2nd ed.). New York: Springer.

Lissitz, R. W. (2009). *The concept of validity*. Charlotte: IAP-Information Age Publishing.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27*(3), 32–42.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*, 91–98. doi: 10.1111/jedm.12003

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement*, *50*, 105–109. doi: 10.1111/jedm.12004

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Charlotte, NC: Information Age Publishers.

Sireci, S. G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement*, 50, 99–104. doi: 10.1111/jedm.12005

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8*, 63–82.

Chapter 4

Collecting validity evidence

Comparable performance standards in arithmetic for different educational tracks in the Netherlands.

Abstract

In this study, validity evidence was gathered to support claims regarding the validity of the decisions made within the reference levels assessment programme that was recently introduced in the Netherlands. It focuses on the IRT-based linking procedures that are crucial for the validity of decisions made within this assessment programme. Data for 80 arithmetic items are collected and analysed to examine whether the assumptions needed for linking procedures are violated and, furthermore, the extent to which this influences the validity of the decisions made within this assessment programme. It is concluded that when common IRT equating procedures are used to compare test scores on different forms, the validity of the test scores is compromised.

Keywords: Validation, Argument-based Approach, Equating, DIF, Comparability

Chapter submitted for publication:
Wools, S., Béguin, A. A., & Eggen, T. J. H. M. (January 2015). Comparable performance standards in arithmetic for different educational tracks in the Netherlands. Submitted for publication in: *Assessment in Education: Principles, Policy & Practice*

Introduction

The Dutch government recently implemented new, fine-grained achievement standards for Dutch language and arithmetic. These achievement standards are called reference levels and are applicable to all students in primary, vocational and secondary education. These reference levels are integrated in an accompanying assessment programme for all levels of education. This programme includes assessments with high-stakes consequences for students. One of the key aspects of this assessment programme is that students from different age groups and different educational tracks need to demonstrate the same proficiency level. Due to differences among age groups, it is not possible to use the same tests for all students; therefore, specific tests are constructed for every age group. To ensure comparability over test forms, IRT-based linking procedures can be used (Kolen & Brennan, 2004). However, these procedures rely heavily on assumptions regarding the data, for example, that the estimated parameters hold for every group of students (Von Davier & Von Davier, 2012). In the Dutch context, since the different groups of students differ significantly, it is questionable whether this assumption will hold. This could cause problems in the validity of decisions made within the assessment programme. These problems relate to the comparability of test forms and, therefore, the comparability of the decisions made about students' proficiency level.

As part of the validation studies performed for this assessment programme, this paper focuses on the IRT-based linking procedures that are crucial for the validity of decisions made within the assessment programme. It purports to examine whether the assumptions needed for these linking procedures are violated and, furthermore, the extent to which this influences the validity of the decisions made within this assessment programme. The results of this study serve as validity evidence regarding the claim that the tests in the assessment programme—which are aimed at different target populations—can be used to make a decision regarding the mastery of the same reference levels. This validity evidence is analysed within the argument-based approach to validity proposed by Kane (2006, 2013).

To gain a good understanding of the assessment programme as well as its methodological challenges and threats to validity, it is necessary to have some background knowledge of the Dutch educational system. This article therefore starts with a summary of the system and an introduction of the newly implemented assessment programme. This first section ends with elaborations

on some methodological challenges and validity issues involved in this assessment programme. The second section describes the validation study on the basis of the magnitude of these challenges. The article concludes by addressing the results within a validity perspective in order to add to an expanding body of knowledge on the validity of the scores obtained in this assessment programme.

The Dutch Educational System and Reference levels

The educational system

The Dutch educational system consists of three levels: primary, secondary and tertiary education. It is characterised by early differentiation of students into different tracks (Scheerens, Ehren, Sleegers, & de Leeuw, 2012). In this article, commonly utilised abbreviations in the Netherlands are used to indicate these different tracks. Both the meaning of these abbreviations and the structure of the educational system are illustrated in Figure 4.1. The figure shows that all students attend primary school (PO) from age four to 12. When students enter secondary school, they are placed into different tracks. These tracks are primarily based on the students' proficiency level. In Figure 4.1, the tracks are displayed in ascending order from left to right in terms of theoretical orientation and complexity. Furthermore, the tracks differ in the type of higher education students prepare for. The three tracks on the left (bb, kb and gt) are pre-vocational in nature and are therefore more practice-oriented (vmbo). The two tracks on the right are more general, theory-oriented pre-academic tracks (havo and vwo). Finally, the tracks differ according to duration of study. The vocational tracks are all four years long, and the pre-academic tracks are five or six years long. Upon completing one of these tracks, students can continue onto secondary vocational education (mbo or tertiary education (hbo or wo). These programmes also have different complexities but are all forms of professional education and require between two and four years, depending on their complexity.

The transition between secondary education and professional education is marked by final examinations. These examinations are specifically constructed with the aim of an appropriate level of proficiency in the specific secondary education track. When students pass these examinations, they are granted access to the connected secondary vocational education (mbo) or tertiary education: higher professional education (hbo) or university education (wo). In

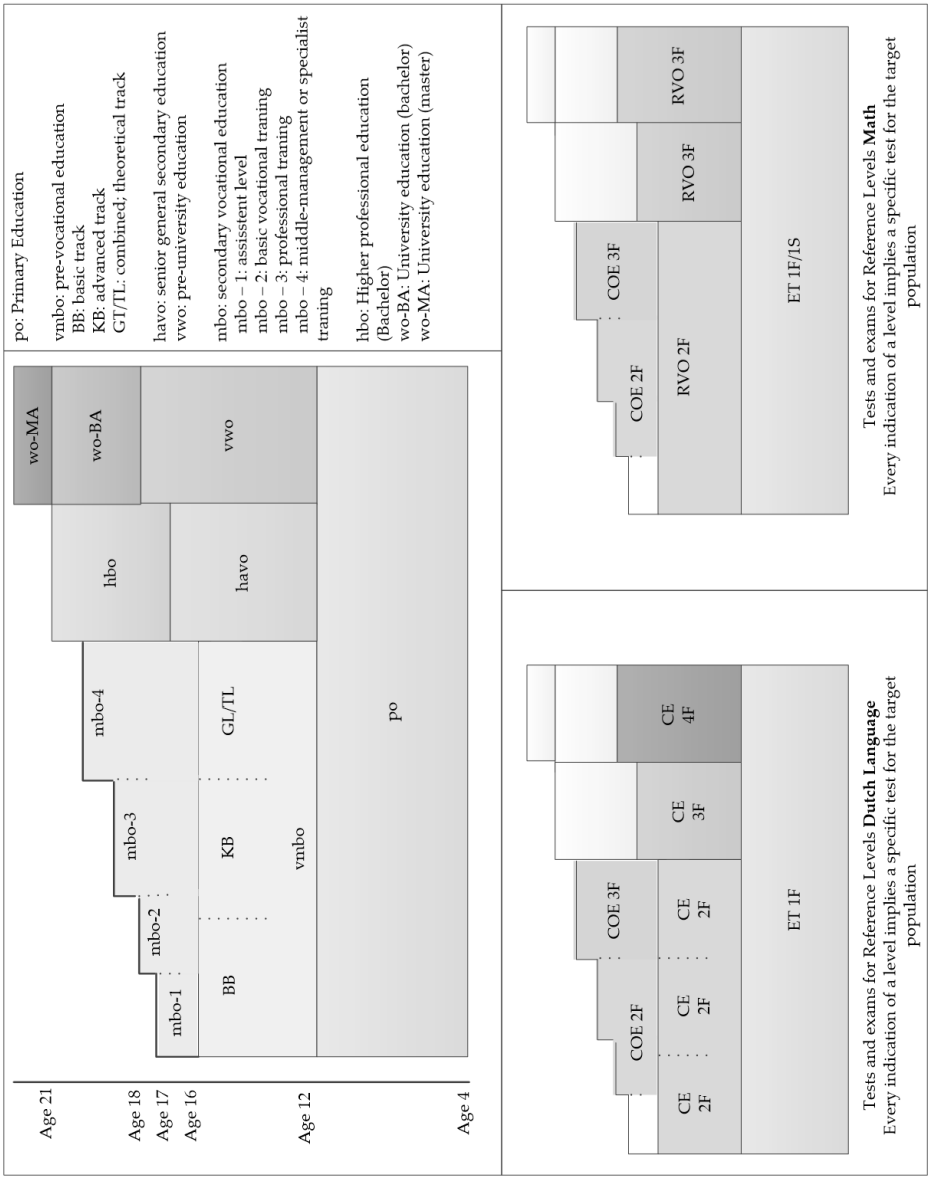


Figure 4.1: Dutch educational system & reference levels

Figure 4.1, these connected tracks are displayed in the same column as the secondary education tracks. In general, it is not necessary to pass an additional admissions test for secondary vocational education, higher professional

education or university. In other words, the final exams also function as an admissions tests to higher education.

Reference levels

In 2010, legislation was implemented to ensure a basic proficiency level in arithmetic and Dutch language for students who are about to enter the labour market or tertiary education. To do so, achievement standards are specified into a description of specific attainment targets: 'reference levels'. Reference levels include fundamental levels of proficiency that students need to achieve at the end of an educational level. For arithmetic, 1F, 1S, 2F and 3F are distinguished whereby 1F, 2F and 3F describe a more functional level in arithmetic, and 1S is a more advanced theoretical level for more advanced students in primary education. For Dutch language, the different levels are 1F, 2F, 3F and 4F. These are displayed in the bottom half of Figure 4.1. This illustration shows that the 1F level in Dutch language and the 1F or 1S level in arithmetic need to be achieved by students at the end of primary school. All students at the end of pre-vocational education need to reach 2F in both arithmetic and Dutch language. This 2F level should be maintained by students who continue to the three least complex levels of secondary vocational education. All students who would like to enter tertiary education, like students in general secondary education and those in the highest track of secondary vocational education, need to reach 3F in arithmetic. For Dutch language, these students need to achieve 3F, or for entry to university, they need to reach 4F.

To assess whether students have achieved the appropriate reference levels, several tests are implemented in the curriculum. Although all tests purport to classify students in masters or non-masters on the reference levels, different tests that fit the target populations are constructed. These tests differ, for example, according to item types and item content. In primary education, for instance, contextual items are aimed at students who are around 12 years old. In secondary vocational education, the items relate to situations that fit the perceptions of 20-year-old students. Also, in some target populations (like primary education), paper-based tests are used, and due to logistical issues, these tests consist primarily of multiple choice items. In other target populations (like secondary vocational education), computer-based tests with open-ended questions are developed. These differences fit the particular testing cultures within the educational tracks.

Implications of this system

With the implementation of the reference levels, only the obligatory levels are defined. In practice, however, more proficient students might reach these levels easily and could aim for the next level. Students in primary education (po), for example, who are likely to go on to pre-university education (vwo) will probably achieve 2F in Dutch language at the end of primary education. At the same time, the 2F level in arithmetic seems very ambitious for the least complex tracks in secondary education, such as the pre-vocational education Basic Track (vmbo BB) and the Advanced Track (vmbo KB).

Despite the different characteristics of the groups of students in the different tracks, they all need to demonstrate that they master a proficiency level above the performance standard defined for the reference levels. So each reference level is defined by a single performance standard that needs to be valid for groups of students with different characteristics. This means that test scores obtained by one population on a particular test must be comparable to those obtained by another population on another test. For example, when students from pre-vocational educational (vmbo) demonstrate a proficiency level of 2F in arithmetic, they need to maintain this level during secondary vocational education (mbo). This implies that the 2F level—measured during the final exams of vmbo—is the same 2F as the one measured in mbo. For Dutch language level 3F, students in a more practice-oriented secondary vocational education track mbo-4) need to demonstrate the same level of ability as those in the senior general secondary education track (havo).

The comparability of the test scores can be obtained by means of statistical techniques that can be used to equate test results acquired through different tests forms (e.g. Kolen & Brennan, 2004). When these forms are aimed at different groups of students, as is the case of the reference level assessment programme, the techniques are referred to as vertical linking procedures (e.g. Carlton, 2011; Harris, 2007; Kolen & Brennan, 2004, pp. 372–418). These procedures make relatively strong assumptions about the properties of the (linking) data and often largely depend on the underlying IRT models.

Validity of the assessment program

The linking of the tests to assess the reference levels and, consequently, the comparability of test scores for different populations is essential in the implementation of the reference levels in the Dutch educational system. Therefore comparability is a crucial aspect of validity research for these high-

stakes tests that students need to pass to receive a diploma in their track. The general view is that validity needs to be supported by presenting validity evidence (AERA, APA, NCME, 1999). Kane (2006, 2013) proposes an argument-based approach to validity. Within this approach, an interpretive argument is identified. This argument aims to make the proposed interpretation and use of assessment scores explicit. Therefore, the inferences and accompanying claims that underlie the proposed interpretation and use need to be specified. By doing so, it becomes apparent which claims and inferences are challenged the most, and these claims can subsequently receive the most attention in validation studies. The evidence collected to support or reject claims is made explicit in the interpretive argument and can be used as part of the validity argument. This validity argument is constructed to summarise all available evidence related to the validity of test scores.

By this reasoning, the comparability of test results within this particular assessment programme should be rigorously evaluated and subject to considerable scrutiny. Therefore, the main validity question that can be posed for this assessment programme is: can we assume that the same reference level is measured in different tests for different populations? This should be reflected in the interpretive argument for this particular assessment programme. Within an assessment programme, the same basic inferences can be identified as those for a single test (Wools, Eggen, & Béguin, submitted): a scoring inference, a generalisation inference, an extrapolation inference and a decision-making inference.

The validity questions posed here are related to claims made within the generalisation and decision-making inferences. Within the generalisation inference, it is claimed that the sample of items used in a test can be used to generalise a score to a hypothetical score on a test domain consisting of all items that could have been presented to students (Wools, Eggen, & Sanders, 2010). One of the major assumptions in this inference is that the model assumptions of the statistical models used to support the generalisation, such as the models used for vertical linking, are met. The decision inference claims that the intended decision is made when the test is used. In the context of this assessment programme, this means that students are correctly classified as masters or non-masters of the reference level. A major possible rebuttal in relation to this latter inference is the fact that the violations of assumptions in statistical models cause an increase in the percentage of incorrectly classified students. In this context, this article raises two research questions, each relating

to one of the two inferences within the validity argument of the assessment programme:

1. Is the order of difficulty of items the same over different tests administered to different populations? (generalisation inference)
2. Is the classification of students according to masters or non-masters consistent over different tests? (decision inference)

The two questions are related to the extent that when the first question is answered positively, the second is also confirmed. The second question, however, can be regarded on its own since it can, in some cases, also be answered positively while the first question is not. Moreover, within the context of reference levels, this latter question seems more fitting for the intended use of the assessment programme. Thus, although it is stronger when the first question is confirmed, the validity of the assessment programme is also supported when only the second question is confirmed. When the answers to these questions provide evidence that supports the claims being made within both inferences, this adds to the body of evidence necessary to validate the reference levels assessment programme.

Methods

To answer the questions posed in this paper, analyses were performed on data collected in a larger comparison study (Wools & Béguin, 2013). This comparison study was conducted as part of the implementation of the assessment programme and included data collection for both Dutch reading and arithmetic among all relevant populations. For clarity purposes, this paper is confined to the available data on 2F arithmetic. These items are administered in most tracks and those tracks in which students are most likely to differ, since data are collected for students aged 12 (po) through 20 (mbo 2/3).
Within the comparison study, so-called reference sets are constructed (Béguin & Wools, 2015). A reference set is a collection of items that is administered within several educational tracks. Furthermore, for each reference set, a performance standard is determined to decide on the number of items within a set that students need to answer correctly in order to demonstrate mastery of the reference level.

Reference set arithmetic 2F

Seven experts selected 80 items to be included in the reference set. To ensure a variety of item types were included in the reference set, both open-ended and multiple choice items were selected. The majority of the items consisted of a functional context that introduces a mathematical problem for students to solve. In some of these cases, students were allowed to use a calculator. The selected items are representative of the attainment targets described in the reference levels. Furthermore, while the items originated from different educational tracks, they were sufficiently general to be administered in other tracks. Twelve items were originally constructed for students aged 12; 41 items were aimed at students aged 16 through 18; and 27 items were constructed for students who are approximately 20 years old and who are ready to begin work. Table 4.1 shows the distribution of items within the set for origin, content domain, question type and permission to use a calculator to solve an item. This table illustrates the diversity of the set and shows the representativeness of the items for the 2F arithmetic domain.

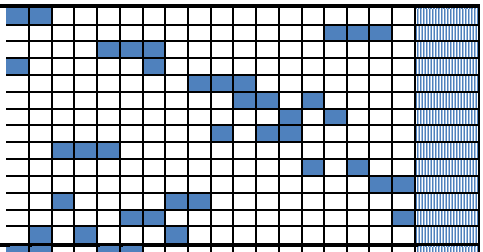
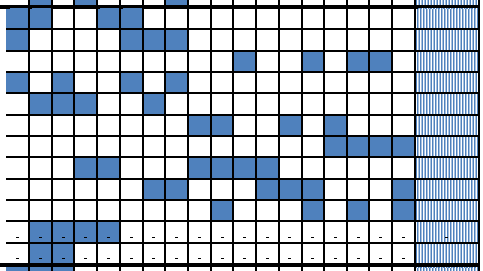
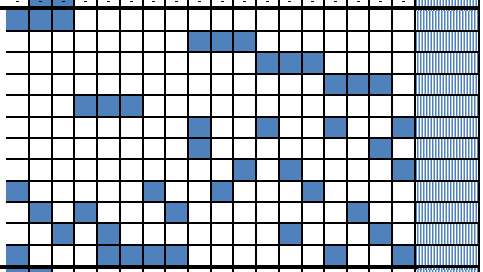
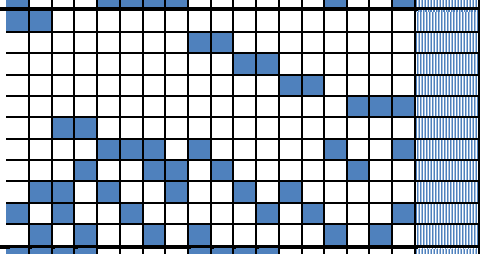
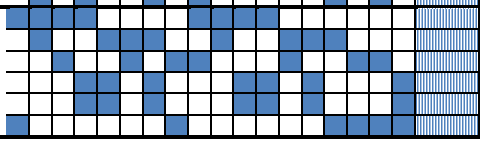
Table 4.1: the composition of the 2F reference set for arithmetic.

Origin		Content domain		Question type		Use of calculator	
12 years	12	numbers	24	multiple choice	34	calculator	44
16 - 18 years	41	measurement	16	open ended	46	non-calculator	36
20 years	27	relations	16				
		ratios	24				
Total	80	Total	80	Total	80	Total	80

Data collection

Once the reference set was constructed, the data were collected. Therefore, the reference set items were administered to students from different educational tracks. Data were collected for students in the last grade of primary education (po) and those in their senior year of pre-vocational education (vmbo), more specifically, in the basic and advanced tracks (vmbo BB/KB) and the combined theoretical track (vmbo GT). The 80 items were also administered in the last two years of senior general secondary education (havo). Finally, data were collected in the second half of the curriculum of levels 2 and 3 or secondary vocational education (mbo 2/3). Depending on the specific programme, this would be after one or two years of education.

The number of items within a reference set is too large for one student to complete during a test session. To solve this, an incomplete design was used to gather data on all the items.

			*
po	Mean N per item	482	
	Minimum N per item	435	
	Maximum N per item	711	
vmbo bb-kb	Mean N per item	544	
	Minimum N per item	302	
	Maximum N per item	815	
vmbo gt	Mean N per item	844	
	Minimum N per item	497	
	Maximum N per item	1112	
havo	Mean N per item	407	
	Minimum N per item	285	
	Maximum N per item	659	
mbo 2/3	Mean N per item	336	
	Minimum N per item	152	
	Maximum N per item	423	

* The data used in this study is collected as part of a larger study, students were presented with more items than displayed here. These additional items are visually represented in the last column.

Figure 4.2: Data collection arithmetic 2F

Data collection for 2F items was part of a larger study that entailed data collection for all the reference levels. Figure 4.2 summarises the data collection for the 2F arithmetic items. Every row represents a test form within a predefined population. A shaded cell represents a cluster of items that is administered in a particular test form. Every cluster of items is administered at least two times within a population. The mean, minimum and maximum number of observations per educational track are displayed in Figure 4.2. The test versions, represented by rows in Figure 4.2, were administered on paper and consisted of two tasks. It took each student 45 minutes to complete a task. When a calculator was permitted, this was the case for all items within a task.

Standard setting

When all the data were collected, a standard setting procedure (Hambleton & Pitoniak, 2006) was performed to define an absolute, content-based standard on the reference set. A group of 15 content experts and teachers participated in this procedure. To determine a performance standard, both an extended Angoff (Hambleton & Plake, 2005) procedure[1] and a bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) were performed. During these procedures, experts decided on the number of items that needed to be answered correctly to pass a reference level. For the final standard setting round, the results of both procedures were evaluated by the experts on the basis of the expected impact on the percentage of mastery in the populations. Upon combining all sources of information, the participants gave their final recommendations for the performance standard. The average of the individual recommendations for the standard was regarded as the outcome of the standard setting procedure and, consequently, as the performance standard for the reference sets. The expert group participating in the standard setting of the 2F items reached a performance standard of 44/45 out of 80 items. This meant that a student, independent of the educational track in which he or she was on, needs to answer at least 45 items of the reference set correctly to show mastery of 2F arithmetic.

[1] Mean Gower's coefficient for rater agreement of the experts performing the Extended Angoff procedure was .79 in the first round and .84 in the second round.

Analysis

To answer the research questions posed in the introduction of this article, analyses are performed on the described data. The methods used to analyse both research questions are described in this section.

Ordering of items

The first question relates to the ordering of items that are administered to different groups of students. To answer this question, item characteristics were studied for the different populations. When the order of items on the basis of difficulty differed among the populations, it is implied that the items do not have the same characteristics for every group. When items hold different characteristics for students from different educational tracks, but with comparable proficiency, this is referred to as differential item functioning (DIF) (Holland & Wainer, 1993). In DIF analysis, there is a difference between the reference group that provides the baseline for the comparison and the focal group. In DIF analysis it is determined whether the focal group deviates from the reference group. Throughout this article, the vmbo GT population serves as a reference group and as the basis for comparison. This is because students in the vmbo GT track are of average proficiency and age compared to the other populations participating in this study.

There are two approaches to DIF analysis: methods that rely on observed score analysis and those that use models from item response theory (IRT) (Camili, 2006). In this paper, DIF in the reference set items was studied using both approaches. First, differences in terms of facilities or p-values over populations were studied. Due to the incomplete design used to collect data, common non-parametric models, such as the Mantel-Haenszel technique (Mantel & Haenszel, 1959), could not be used. Therefore, p-values were calculated using empirical data. Subsequently, the ordering of p-values was compared for the different student groups. The items were ordered according to their p-values in the reference group (GT). The p-values of the items in the consecutive focus groups were plotted against the p-values of the GT reference group. When the ordering of item difficulty differs, this is made visible by the deviations from the diagonal. To the extent that these differences are small, they can be attributed to sampling error. However, larger differences in item ordering can indicate DIF.

An IRT-based evaluation of DIF was used alongside the analyses of the p-values. Two parameter logistic (2PL) IRT models were estimated for both populations separately using Bilog-MG software (Zimowski, Muraki, Mislevy, & Bock, 1996). The data were analysed using multiple forms in an incomplete

design and a single group for all the forms within a population. To make comparisons between the populations possible, the IRT parameters of the items in the consecutive focal groups were then linked to the same scale as the items in the GT reference group. This was done by means of a Stocking Lord transformation (Stocking & Lord, 1983) using the ST programme (Hanson & Zeng, 1995; Hanson, Zeng, & Cui, 2004). In this procedure, the focal group scale was transformed using a linear transformation of the reference group scale. This was done so that the difference between the item characteristic curves (ICCs) of the items in both scales are minimised.

The amount of DIF was then quantified in two different ways. Firstly, a residual analysis was carried out between the transformed parameters of the focal group and the parameters of the same items in the reference group. When items do not show any DIF, the transformed parameters are identical to the original parameters. If there is DIF, we can evaluate the difference between the two scales in terms of the difference in the probability of answering an item correct for two students with the same proficiency: an average GT student and a student from another population with the same proficiency as the average GT student.

Secondly, the difference between the ICCs can also be used as an indication of the amount of DIF. This type of DIF measure, based on the area between the ICCs, was introduced by Raju (1988) and further developed by Raju, Van der Linden and Fleer (1995). The latter applied a weighting of the area between the ICCs based on the proficiency distribution of the population of students. To weigh the proficiency, the distribution of the focal group was used. Both signed and unsigned area statistics were calculated (Camilli, 2006 pp. 236–237). In both cases, the larger the area, the more DIF there was.

Classification of students

The second research question concerns the consistency of the classification of students as masters and non-masters. Differences in the classification of students can occur due to DIF in the anchor. For example, when the composition of the anchor used to transfer the performance standard leads to different cut scores. To get a better understanding of the impact on classification consistency, the impact of DIF in the anchor is shown in the context of an operational test.

To exemplify the effect of DIF in the anchor on transferring the performance standard to different tests, the expected scores of an average student on an operational test are calculated. The operational test used in this paper is aimed

at students in the GT track who need to demonstrate ability at the 2F level (RVO 2F, Figure 4.1). The parameters for this 60-item operational test were calculated for the GT population on the basis of a 2013 test administration. To estimate the expected score of the average student from other populations, different subsets of items, or anchors, were used. The differences in the expected scores for the average student, depending on the items in the anchor, show the impact of the DIF items in an anchor. This procedure was repeated for all relevant populations.
Subsequently, the impact of different items in the anchor, which are used to transfer the performance standard is evaluated in terms of classification consistency. The expected scores are therefore evaluated against the cut score of the operational test. The latter gives a better understanding of the consequences of using an anchor with DIF items.

Results

The results section is structured according to the two research questions. The results used to answer the first question are presented in 'Ordering of items' while those corresponding to the second research question are discussed in 'Classification consistency'.

Ordering of items

In the methods section, two approaches are described to study DIF in the reference set items. The first focuses on observed score analysis while the second focuses on the application of IRT.

Observed score-based evaluation of DIF

First, the ordering of the facilities of items (p-values) within the 2F reference set in four different populations is displayed in Figure 4.3. The GT population served as the reference group, and the po, BBKB, havo and mbo 2/3 populations served as focal groups for the comparison. The dotted line indicates the absolute difference in the p-values resulting from the difference in the proficiency of the comparison population. For example, in the upper left corner, the po population was less proficient than the GT population; therefore, the mean p-value was lower for the po population than for the GT population. The dots concentrated around the line represent items that are ordered from easy to difficult for the GT population. When the distance between a data point and the

dotted line is larger, this means that an item is relatively easy or difficult for the focal group. In Figure 4.3, the comparison between the GT and po populations shows the largest differences in item ordering. Within the havo plot, the items are grouped together in the upper right corner. This ceiling effect is due to the fact that the items are relatively easy for these students. From the mbo 2/3 plot, we can see that this population is very similar to the GT population, both in terms of ability level (the two dotted lines are almost identical) and item ordering – most data points are grouped around the lines.

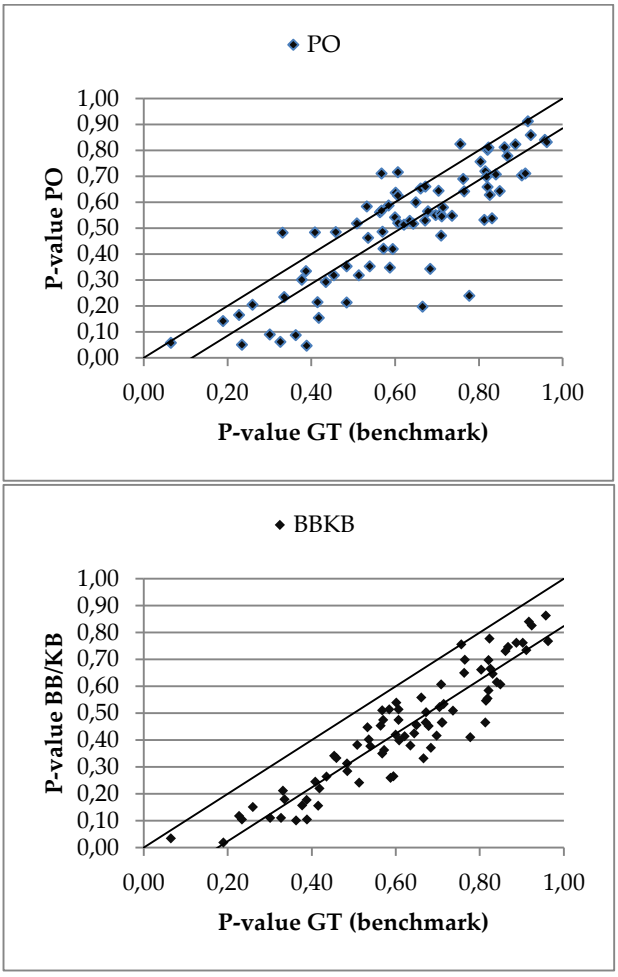


Figure 4.3A: 2F- ordering of P-values (PO & BBKB)

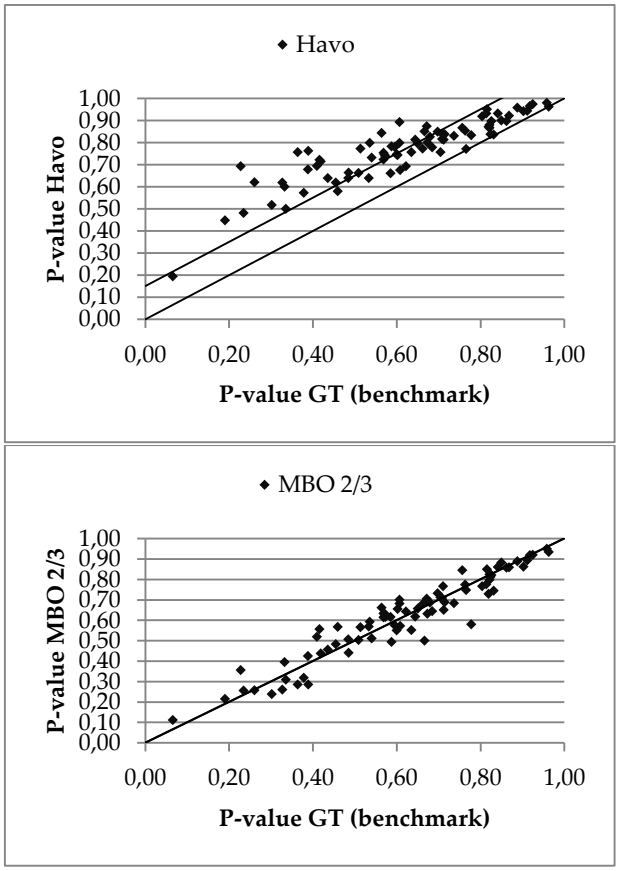


Figure 4.3B: 2F- ordering of P-values (havo, mbo 2/3)

IRT based evaluation of DIF

In this study, IRT models were estimated for all populations separately. The IRT parameters of the items in the focal group were then linked to the same scale as the items in the reference group (GT) using a Stocking Lord transformation (Stocking & Lord, 1983). Appendix I includes graphs indicating the differences in parameters for all focal groups and the reference group.

When the items were linked to the same scale, the residuals of the linking were evaluated in terms of the probability that an average student answers an item correctly (expected p-value). It was found that for the average po (primary education) student, the differences between the parameters for po and the parameters for GT (pre vocational education) result in 31 items with an absolute difference in an expected p-value of more than 0.10. Of these items, 10 had a

higher absolute difference than 0.25. The average effect reaches 0 by definition, and the standard deviation of the effect is 0.14. The results of the comparison for other populations are shown in Table 4.2. These effects are smaller, but still 10 or more out of 80 items show an absolute difference in expected p-value of .10 or more.

Table 4.2: number of items with absolute differences in p-value of more than 0.10.

	po	**BBKB**	**havo**	**mbo 2/3**
difference > .25	10	0	1	0
difference > .10 < .25	21	14	10	11
difference < .10	49	66	69	69
Mean	-0.003	-0.001	-0.001	-0.001
Standard deviation	0.146	0.082	0.080	0.068

Table 4.3 gives a summary of the DIF statistics based on the weighted difference between the ICCs for the reference group (vmbo-GT) and the focal groups. In BBKB, six items had a weighted signed difference of more than 0.15, and as many as 39 items had a weighted signed difference of 0.05. When unsigned differences were used, more than half (= 44) of the items showed an average weighted difference of at least 0.05. This last result is in line with the average weighted signed difference of 0.06. The results for mbo-2 and Havo were, to a large extent, similar to those of vmbo-BBKB.

Table 4.3: Summary of the ICC based statistics

	po		**BBKB**	
	signed	**unsigned**	**signed**	**unsigned**
number > abs(0.15)	24	13	6	6
number > abs(0.10)	41	25	12	12
number > abs(0.05)	56	46	39	44
average	-0.028	0.086	0.00	0.06

	havo		**mbo 2/3**	
	signed	**unsigned**	**signed**	**unsigned**
number > abs(0.15)	7	7	4	4
number > abs(0.10)	11	11	9	9
number > abs(0.05)	30	35	25	29
average	0.00	0.06	0.00	0.05

Classification consistency

The expected score of an average student from the different populations was estimated for a 60-item test for students in pre-vocational education. This expected score was calculated for several different anchors. The anchors were chosen on the basis of item characteristics. For example, one anchor consisted of all items that were presented with a calculator while another anchor consisted of all items for which a calculator was not allowed. The results of this analysis are presented in Table 4.4.

This table shows differences of up to 7.9 score points in the expected score of an average po student when using an anchor with only items that could be solved using a calculator as opposed to an anchor with only non-calculator items. Other differences were smaller but still present, for example, an anchor with items of one content domain would lead to a difference of 3.9 score points in the expected score of a mbo 2/3 student with an average ability.

Table 4.4: expected score for average student of population on 60 item operational test based on different anchors

	po	**BBKB**	**havo**	**mbo 2/3**
All items (baseline)	27.55	22.24	44.66	34.92
Origin	**po**	**BBKB**	**havo**	**mbo 2/3**
po	30.21	24.06	45.39	35.10
vo	27.13	22.27	44.10	35.27
mbo	27.06	21.44	45.30	34.33
Content domain	**po**	**BBKB**	**havo**	**mbo 2/3**
Numbers	30.83	23.09	44.54	36.60
Relations	24.88	20.84	45.83	32.70
Ratios	27.55	22.9	45.07	35.15
Measurement	25.69	21.45	43.20	34.42
Question type	**po**	**BBKB**	**havo**	**mbo 2/3**
Open-ended	27.09	21.55	44.4	34.68
Multiple choice	28.30	23.18	44.96	35.37
Use of calculator	**po**	**BBKB**	**havo**	**mbo 2/3**
Calculator	24.29	20.18	45.19	34.18
Non-calculator	32.18	24.86	43.86	36.02

To evaluate the influence of different anchors used to transfer the performance standard in terms of classification decisions, the same anchors were used as in the evaluation of the expected score. After the linking using the different anchors, the proportion of the population scoring above the cut score of 33/34 was calculated. The results of this analysis are given in Table 4.5.

Table 4.5: percentages of students that would pass the operational test dependent on different anchors

	po	**BBKB**	**havo**	**mbo 2/3**
All items (baseline)	0.27	0.13	0.87	0.56
Origin	**PO**	**BBKB**	**Havo**	**MBO 2/3**
PO	0.41	0.14	0.87	0.57
VO	0.26	0.14	0.86	0.57
MBO	0.23	0.13	0.89	0.55
Content domain	**po**	**BBKB**	**havo**	**mbo 2/3**
Numbers	0.43	0.13	0.87	0.59
Relations	0.16	0.10	0.90	0.52
Ratios	0.24	0.13	0.87	0.57
Measurement	0.19	0.13	0.85	0.55
Question type	**po**	**BBKB**	**havo**	**mbo 2/3**
Open-ended	0.24	0.13	0.88	0.56
Multiple choice	0.35	0.13	0.86	0.57
Use of calculator	**po**	**BBKB**	**havo**	**mbo 2/3**
Calculator	0.15	0.09	0.88	0.55
Non-calculator	0.45	0.19	0.86	0.57

The percentages presented in Table 4.5 show the expected proportion of students that would pass the operational test. The proportion of students that would pass the test when all items are included could be interpreted as the true percentage; all deviations from this proportion due to DIF in the anchor can be seen as percentage of misclassification. For example, using an anchor with calculator items means that an additional four percent of the BBKB students would be wrongly identified as non-masters.

Conclusion

In this study, validity evidence was gathered to support claims regarding the validity of the decisions made within the reference levels assessment programme that was recently introduced in the Netherlands. The validity evidence focused on the question of whether the tests in the assessment programme, aimed at different target populations, can be used to make a decision regarding the mastery of the same reference levels. More specifically, two research questions were posed in relation to claims being made within the generalisation inference and the decision-making inference specified within the argument-based approach of Kane (2006, 2013).
In this section, the research questions are answered in relation to the inferences of the interpretive argument. Since the results are part of the body of knowledge relating to the validity of the assessment programme, both inferences are described. Furthermore, the earlier presented validity evidence is evaluated in light of these inferences.

Generalization inference

In general, a generalisation inference claims (Wools, Eggen, & Sanders, 2010) that scores can be generalised into a hypothetical score on a test domain. A test domain consists of all the possible tasks that could have been presented. It is assumed that the tasks that are presented to students can be considered as a representative sample of the test domain. Furthermore, when statistical models are used to support the generalisations, the model assumptions must be met. For example, when items within a test domain are calibrated using an IRT model, item parameters are assumed to be equal for all students.
Within the reference level assessment programme, scores are obtained through different test forms. These forms are designed to match a specific target population. Therefore, the assumption that all test forms consist of a representative sample of tasks from the complete test domain is challenged. However, when it is possible to present evidence that the scores obtained through different test forms are comparable, this could counterbalance the claim that test forms are not fully representative from a content perspective. Therefore, in this article, a question was raised regarding the ordering of items administered in different populations. When this ordering is stable for all populations, this could be regarded as evidence of a comparable measure on the basis of different test versions.

The results presented in this article indicate differences in item characteristics for different populations. These differences were found when we examined the p-values but also in the more formal DIF analyses. We were not able to identify variables that could explain what items were relatively easy or difficult for a specific population. Thus, we cannot predict the content elements that need to be emphasised within test forms to enable comparisons. We conclude that assumptions relating to the comparability of test versions are not met within the reference level assessment programme. Therefore, when common IRT equating procedures are used to compare test scores on different forms, the validity of the test scores is compromised.

Decision inference

A decision inference is related to the justification of decisions regarding students that are made based upon test performances (Wools, Eggen, & Béguin, submitted). It is necessary to evidence that the intended decision is made, for example, that the decision is based on performance indicators that match the intended behaviour. Another example is that when classification decisions are made, students are consistently classified as either masters or non-masters, independent of the characteristics of a test version.

In the reference level assessment programme, one performance standard is transferred to tests constructed for different populations. It is assumed that this performance standard represents the same level of proficiency within all tests. However, for this assumption to be accepted, the effects of the violation of the statistical assumptions underlying vertical linking methods have to be small. This is so because when DIF items are included in the anchor used to transfer the performance standard, students might be wrongly classified as masters or non-masters. More specifically, students may be classified as masters when taking a test form that includes items that are relatively easy while the same students may be classified as non-masters when a test form includes items that are relatively difficult.

This article shows that the effect of DIF on the anchor used to transfer the performance standard to different tests could lead to large differences in expected scores for an average student on an operational test. When these differences were evaluated against the cut score of this test, the percentage of students that would have been wrongly classified became apparent. Especially in the context of high-stakes exams, this would lead to an undesirable situation and would invalidate the results of the assessment programme.

Discussion

The results of this study and the conclusions regarding the validity of the assessment programme have led us to conclude that traditional linking and equating techniques (Kolen & Brennan, 2004) are not suited to this particular situation. Therefore, it is necessary to implement another method that enables comparisons of students from different populations. Given the intended interpretation and use of the reference level assessment programme, this method should make comparisons of students possible and should lead to high classification consistency for the master/non-master decision. A more promising procedure is that in which comparisons over different educational tracks are not based on a common IRT scale but on a number correct or common items, such as a reference set. This procedure (Béguin & Wools, 2015) creates a more stable basis for comparison that fits the intended purpose of the assessment programme.

Whether this procedure could solve the validity issues raised in the conclusion of this paper should be evaluated by means of the data collected in the comparison study used in this article. However, just as the results presented in this paper, the interpretation of such an evaluation will be complicated by the study's data limitations. These limitations include, for example, that data were collected in a low-stakes testing condition while the statistical inferences based on the data were in relation to a high-stakes testing condition. Student performance in low-stakes conditions could be different from normal or high-stakes conditions (Wise & DeMars, 2005), and this would potentially affect the results of linking and equating procedures (Keizer-Mittelhaëuser, Béguin, & Sijtsma, 2015). Furthermore, the current results could change over time since this study was conducted as part of the implementation of the new educational system. Students were not as prepared as they might be in the future, and therefore, their performance was most likely underestimated. The effect of this underestimation might not necessarily be the same for students from different school types. In this way, relative positions and the comparison of test forms could change over time.

It should also be noted that for the validation of an assessment programme like the reference level assessment programme, further validation studies are necessary. Moreover, all results, including those that support and reject specific claims, should be combined in one coherent validity argument (Wools & Eggen, submitted). Therefore, to gain a full understanding of the validity of this assessment programme the results of this study should not be interpreted in isolation from those of other validation studies.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington: American Psychological Association.

Béguin, A. A., & Wools, S. (2015). Vertical comparison using reference sets. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & W.-C. Wang (Eds.), *New developments in quantitative psychology: Presentations from the 78th annual Psychometric Society meeting* (pp. 195-211). New York: Springer. doi: 10.1007/978-3-319-07503-7_12

Camili, G. (2006). Test fairness. In R. L. Brennan (ed.), *Educational measurement* (4th ed., pp. 221–256). Westport: American Council on Education and Praeger Publishers.

Carlton, J. E. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York: Springer.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*, 41–55.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport: American Council on Education and Praeger Publishers.

Hanson, B. A., & Zeng, L. (1995). ST: A computer program for IRT scale transformation.

Hanson, B. A., Zeng, L., & Cui, Z. (2004). ST: A computer program for IRT score transformation [Computer software]. Retrieved from http://www.education.uiowa.edu/casma/computer_programs

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–252). New York: Springer.

Holland P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale. NJ: Lawrence Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Keizer-Mittelhaëuser, M., Béguin, A. A., & Sijtsma, K. (2015). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & W.-C. Wang (Eds.), *New developments in quantitative psychology: Presentations from the 78th annual Psychometric Society meeting* (pp. 181-193). New York: Springer. doi: 10.1007/978-3-319-07503-7_11.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating* (2nd ed.). New York: Springer.

Scheerens, J., Ehren, M., Sleegers, P., & De Leeuw, R. (2012). *OECD review on evaluation and assessment frameworks for improving school outcomes. Country background report for the Netherlands.* Brussels: OECD. Retrieved from http://www.oecd.org/edu/school/NLD_CBR_Evaluation_and_Assessment.pdf

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 25,* 15.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19,* 353–368.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Von Davier, M., & Von Davier, A. A. (2012). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 225–242). New York: Springer.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wools, S. & Béguin, A.A. (2013). *Onderzoek referentiesets rekenen: datafile en handleiding* [Research project reference sets arithmetic: data file and manual] Retrieved from http://toetswijzer.kennisnet.nl/html/referentiesets_openbaar/default.shtm.

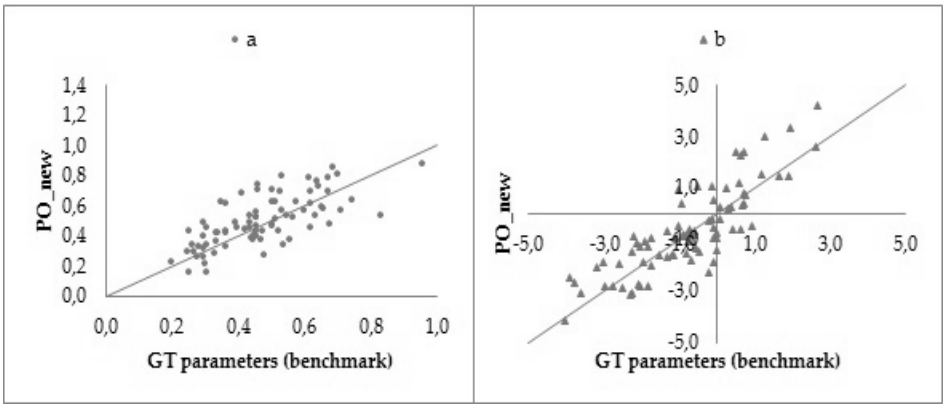
Wools, S., Eggen, T. J. H. M , & Sanders, P. F. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8*, 63–82.

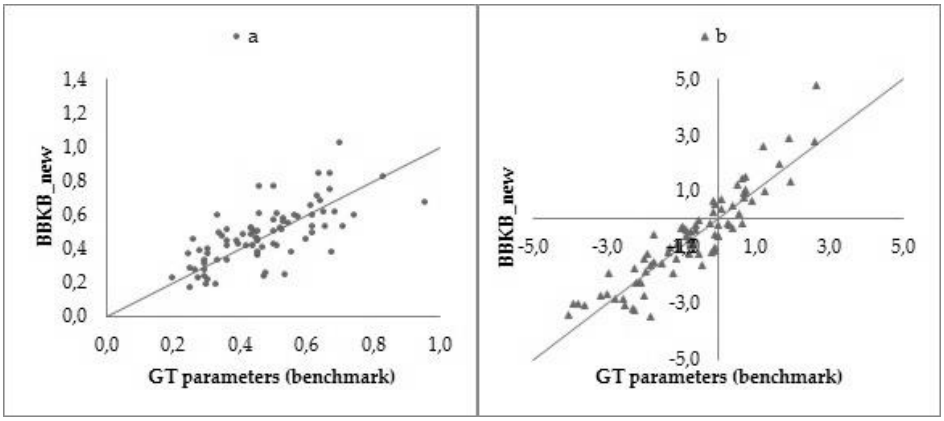
Wools, S., Eggen, T. J. H. M, & Béguin, A. A. (submitted). Constructing validity arguments for test combinations.

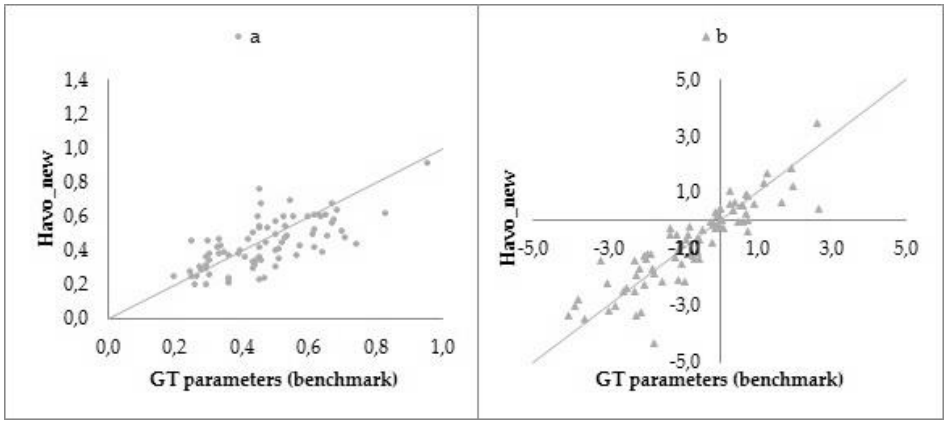
Wools, S., Den Otter, D., & Eggen, T. J. H. M. (submitted). Systematic literature review of validation studies on assessments.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.

Appendix I


a
PO_new
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
GT parameters (benchmark)
b
PO_new
5,0
3,0
1,0
-5,0
-3,0
1,0
3,0
5,0
-3,0
-5,0
GT parameters (benchmark)


a
BBKB_new
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
GT parameters (benchmark)
b
BBKB_new
5,0
3,0
1,0
-5,0
-3,0
1,0
3,0
5,0
-3,0
-5,0
GT parameters (benchmark)


a
Havo_new
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
GT parameters (benchmark)
b
Havo_new
5,0
3,0
1,0
-5,0
-3,0
-1,0
1,0
3,0
5,0
-3,0
-5,0
GT parameters (benchmark)

Chapter 5

Systematic literature review of validation studies on assessments

One of the major theoretical frameworks in validity theory is the argument-based approach advanced by Kane (2006, 2013). This approach suggests that validation is guided by means of an interpretive argument that indicates which inferences need the most attention. In this study, 178 articles on validation efforts in educational assessment are analyzed. All sources of evidence presented by the researchers are classified within a theoretical model of the interpretive argument. It was hypothesized that depending on the intended use of the tests, authors would present different sources of validity evidence. The results show that this is the case for assessments constructed for selection purposes. The validity of these tests is more frequently supported with evidence relating to the possibility of accurately predicting future behavior. Tests with other intended uses did not differ in terms of the sources of evidence they provided. The results also show that the majority of the articles presented only one or two sources of evidence instead of evidence supporting a full validity argument.

Keywords: Assessment, Literature review, Validity, Validity arguments

Chapter submitted for publication (March 2015):
Wools, S., & Den Otter, D., & Eggen, T. (2015). A systematic literature review of validation studies on educational assessments. *Educational Assessment.*

Introduction

In the field of educational assessment, validity is often seen as one of the most important aspects of achievement tests and assessments. Although the concept of validity is continuously under debate (Lissitz, 2009), it is commonly agreed that test scores should be valid and reliable. Especially when high-stake decisions are made on the basis of test results, it is necessary to conduct an extensive study on their validity. The 1999 *Standards for Educational and Psychological Testing* perceives validation as "developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA, APA, NCME, 1999, p. 9). Kane (2004) advances an argument-based approach to validation in order to guide the process of building such a validity argument. This approach focuses on the perspective that the kind of validity evidence required depends on the proposed interpretation of test scores. Following this perspective, the amount and sources of evidence needed might differ according to the different uses of tests. This study investigates whether the amount as well as the sources of validity evidence that researchers present in their documentations of validation studies differ on the basis of the intended interpretation and use of test scores.

Theoretical framework

Due to extensive research on validity, the concept has changed over time (Lissitz, 2009). To summarize the course of the debate on this concept, Kane (2006) cites three aspects of validity that emerged from the widely accepted model of construct validity introduced by Cronbach and Meehl (1955) as general principles of validation. The first principle refers to the increased need to specify the proposed interpretation of test scores. The second principle concerns the premise that evidencing construct validity involves extensive research while the third principle relates to the need to challenge proposed and competing interpretations. These general principles are all accounted for in theories on validity and approaches to validation, for example, in Messick's (1989, p. 13) definition of validity: "…an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (italics in original).

Messick's conceptualization of validity has since been translated by practitioners into a validation practice with the aim of presenting as much validity evidence as possible. From this interpretation, the validity of test scores has been supported by combining countless sources of validity evidence that are either content-related, criterion-related, or construct-related. To lessen the burden on practitioners of collecting different kinds of evidence and, at the same time, ensuring that the most relevant sources of evidence are presented, the argument-based approach to validation was further developed. According to Kane (2013, pp. 8–9):

> The argument-based approach was intended to avoid the need for a fully developed, formal theory required by the strong program of construct validity, and at the same time to avoid the open-endedness and ambiguity of the weak form of construct validity in which any data on any relationship involving the attribute being assessed can be considered grist for the mill. (Bachman, 2005; Cronbach, 1988; Haertel, 1999; Kane, 1992)

The argument-based approach to validation, as proposed by Kane (2006), includes building two arguments: an interpretive argument and a validity argument. The interpretive argument states which inferences and assumptions underlie the intended use and interpretation of test scores. This argument is modelled as "a train of thought," as presented in Figure 5.1.

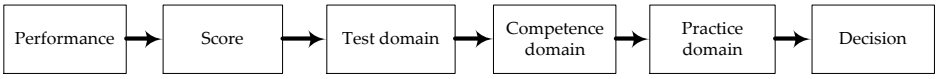


Figure 5.1: Chain of inferences within interpretive argument (Wools, Eggen, Sanders, 2010).

Within this interpretive argument, there is a need for specific inferences in order to make decisions on the basis of observed performance in an assessment situation. The first inference (scoring) relates to students' performances on tasks that are translated into numerical scores. These observed scores are seen as a generalizable instance of the test domain scores (generalization). A test domain represents all possible tasks that could be presented to students. The test domain scores are subsequently extrapolated (extrapolation I) to scores on a competence domain, which entails an operationalization of the competence or construct that is being measured. Within the subsequent inference (extrapolation II), the scores are extrapolated toward a practice domain, that is, a real-life situation that students can encounter in their future (professional)

lives. Building on this final extrapolation, the final inference (decision) can lead to a decision on the students' level on the competence of interest.
When the assessment is fully developed and the interpretive argument is specified, a critical evaluation of the claims in this argument should be performed. This critical evaluation consists of both appraising currently defined inferences and assumptions and rejecting competing interpretations. To do this, both analytical and empirical evidence should be used. The analytical evidence could entail, for example, conceptual analyses and judgments on relationships between the test domain, the competence domain, and the practice domain. Most of the analytical evidence would have already been generated during the development stage of the assessment. The empirical evidence consists, for example, of evidence that relates to the reliability of an assessment, the structure of the construct, or relations with other measures of the construct of interest. This kind of evidence is gathered in so-called validation studies, which are designed to answer specific research questions derived from the need for specific empirical evidence. The results of these studies and the analytical evidence are combined and integrated into a validity argument.
The evidence gathered to evaluate the claims in the interpretive argument can be classified by following the five inferences, which are represented by arrows in Figure 5.1. According to Kane (2009, p. 49):

> …some statements in the literature can be interpreted as saying that adequate validation requires that every possible kind of validity evidence be developed for validation to be complete.… This shotgun approach is clearly unwieldy, and in its extreme form, it makes validation impossible.

Therefore, within the argument-based approach, it is argued that inferences that seem weak or that are of great interest to the intended interpretation and use of tests require more evidence than others. Although evidence is needed for every inference, the weight placed on different inferences depends on the assessment that is being validated. We hypothesize that different inferences are emphasized and supported with evidence for tests with different intended interpretations and uses. For individual students, Schmeiser and Welch (2006) list five purposes for assessments. Alongside the usages for individual students, assessments are also used to gather information about teachers (e.g., accountability) or groups of students (e.g., PISA). These alternative usages are taken together in this study and are not considered independently.
The purposes distinguished by Schmeiser and Welch (2006) represent an intended use and interpretation and hold their own inferences. Table 5.1 shows

these purposes and indicates which inferences we expect to be emphasized in the interpretive argument and are thus most likely to attract attention during the validation process. It was not always possible to indicate a single inference that needed to be most evidenced; for those usages, two inferences are indicated. Note that emphasis on these inference are is to expected, but that it is very well possible that a specific test is developed in a way that another inference should be emphasized than is listed in this table.

Table 5.1: Hypothesis on different evidences for different intended uses

	Placement	Diagnosis	Selection	Classification	Progress
1. Scoring					X
2. Generalization					X
3. Extrapolation I	X	X			
4. Extrapolation II			X		
5. Decision	X			X	

Tests developed for placement purposes are used for course placement and during the counseling of students. It is important that these tests consist of items that elucidate knowledge needed to enter a specific course. Therefore, we expect evidence on the construction of items to demonstrate that items aim at the intended performance (first extrapolation inference). For placement purposes, it is also important that students are correctly classified into categories for placement by means of correct cut scores (decision inference). When a tests is developed for diagnostic purposes, the most important aspect involves its ability to make claims about mastery across various domains. Items should therefore entail the right mastery level and, furthermore, should represent the construct of interest in great detail (first extrapolation inference). Tests and assessments used for selection purposes require evidence that at least supports the claim that we are able to extrapolate scores to a practice domain (second extrapolation inference); that is, selection tests are usually used to predict possible behavior outside the particular testing context. A specific example of a selection test is that used for admission to study programs. For these admissions tests, the greatest validity claim relates to the prediction of future study success; therefore, the second extrapolation inference is most relevant. For classification tests, which are often used for certification purposes, the decision inference is most important since the main question in these tests is: are you able to distinguish between students based on a criterion? Finally, for tests aimed at assessing progress, the intended use could be an assessment of whether students achieved learning goals and whether they show growth over time. Often, these tests are curriculum-based measures with open-ended

question formats that result in claims about the quality of scoring and consistency over raters (scoring inference). Moreover, item sampling, the representativeness of the test, and the measurement of change over time are of importance here, all reflected in the generalization inference.
This study aims to describe the current validation practice in the field of educational assessment. More specifically, it purports to identify sources of validity evidence that researchers present when validating a test or assessment. The study also aims to investigate whether these sources of evidence vary for different types of assessments.

Methods

A systematic literature review was performed in order to study current validation practices. The review focused on answering three main questions.

1. What specific sources of validity evidence do researchers report in articles on validation studies on educational assessments?
2. Can we structure these sources using the inferences specified within the general form of an interpretive argument?
3. Are there differences between the sources of validity evidence gathered and reported for various kinds of interpretations and uses of tests?

Data sources

This article specifically describes validation efforts on assessments in education intended to assess skills, competences, and knowledge learned in educational programs. Articles describing tests intending to measure psychological constructs, such as motivation, social behavior, or self-efficacy, are therefore not included in this study. Furthermore, only articles in which validation was the main purpose were selected for this study. This was aimed at ensuring that all researchers were publishing within the same theoretical rationale of providing validity evidence. Differences in sources of validity evidence could therefore be attributed to another operationalization of validity or validation and not because of another main purpose of the articles.
Since this study limits itself to the evidence presented by researchers on validation efforts for educational assessments, the decision was made to include peer-reviewed articles published between January 1990 and January 2015, indexed in two databases containing journals relevant to educational sciences: ERIC and PsychInfo. To obtain the correct search string, several searches were

performed. In the end, the decision was made to choose the combination of search words that provided us with a broad sample and that still had many relevant sources. This was evaluated by combining different terms and sorting on the basis of "relevance" in the databases. The following search string appeared to be both broad, manageable, and sufficiently specific and was therefore used to obtain articles for this study (January 2015):
(TI "Validity OR TI Validation) AND (Educational Measurement) AND (Assessment OR Test), Limiters – Publication year 1990 – 2014 and peer-reviewed articles.
This search provided us with articles whose titles included the terms validity or validation. Furthermore, only articles referring to educational measurement and explicitly mentioning an assessment or test were included. This initial search resulted in a list of 1,776 unique articles in the databases. From this set of articles, those not written in English were omitted. To narrow the set down even further, only articles from journals relating to the field of education were included.
A total of 1,083 articles were retained for a preliminary screening of titles. Based on title selection, the abstracts of 353 articles were read, and a sample of 215 articles was selected. Full-text versions of these articles were collected for further analysis. Unfortunately, four articles were not available in full text. This left us with a final sample of 211 articles. When the content of these articles was coded, another 33 articles appeared not to be in the scope of this research project. A total of 178 articles were thus included for analysis in the study.
We acknowledge that the choices made in the search process result in a sample of articles that is not exhaustive. However, the sample consists of a selection of articles that present validation efforts for educational assessments. Given the diversity of the journals (n=71) and the wide range of tests described in the articles, we remain confident that the articles form a representative sample of all available articles on this topic.

Data extraction

To ensure that every article was processed in the same manner, a data extraction form was used (Pettigrew & Roberts, 2006). The form was adjusted once during the course of the study to ensure that all relevant aspects were taken into account. The form included three categories: (1) general characteristics of the article, (2) characteristics of the test or assessment, and (3) sources of evidence. The form is included in Appendix I.

1. Characteristics of the articles

In this section of the form, general aspects of the article were collected, such as the title, author/s, journal, and year of publication. Furthermore, the research goal, as described by the author/s of an article, was noted.

2. Characteristics of the test

Domain

Six categories were distinguished within the variable "domain." "Language" was used for tests consisting of measures of constructs like reading, writing, and speaking. "Medical" was used to classify articles describing validation efforts in medical education contexts. "Combination" was used for tests intending to measure multiple constructs, like a school readiness test or admissions tests that include a language section as well as a mathematics section. "Mathematics" and "Science" were used to indicate assessments regarding these specific subjects. The category "others" was used for tests that did not fit into the other categories, for example, computer skills, teacher competence, and visual art.

Educational Level

Four categories were used to describe the target population of validated tests. The selected articles described the validation of tests aimed at pre-schoolers, students in primary education, students in secondary school, or students attending forms of higher education (college or university). When the target population of tests was not described in an article, it was indicated as "unknown."

Intended use

The categories used to describe the intended use of an assessment were derived from the test purposes used, as proposed by Schmeiser and Welch (2006). These five purposes, described earlier in this article, are: Placement, Diagnosis, Selection, Classification, and Progress. When articles described tests intended for usage at a system level (such as tests for accountability purposes) or tests designed for research purposes, these were labeled as "other." When articles specified no intended use of a validated test, they were labeled as "unknown."

3. Sources of evidence

The second part of the form concerned the sources of validity evidence presented in the articles. These sources were derived from guidelines for validity evidence (AERA, et al., 1999; Kane, 2006; Llosa, 2008; Nitko & Brookhart, 2007). For the purpose of this study, the sources of evidence were classified into the five inferences shown in Figure 5.1. When the selected articles presented a source of evidence, it was indicated on the form. Table 5.2 shows the sources of evidence and classifications within the inferences.

Table 5.2: Sources of evidence

Inference	Evidence
1. Scoring	Scoring is correct
	*Rating scheme available
	Rater agreement
	Inter-rater reliability
	Intra-rater reliability
	Other
2. Generalization	Reliability coefficient
	Generalizability coefficient
	*Test blueprint
	Other
3. Extrapolation I	Construct underrepresentation
	Construct irrelevant variance
	External criterion (other test)
	Theoretical model of construct
	Factor analysis
	IRT analysis (e.g., calibration)
	*Authenticity
	*Cognitive labs
	Other
4. Extrapolation II	*Critical tasks
	*Stakeholders
	External criterion (prediction)
	Other
5. Decision	Norm groups
	*Standard setting
	Cut score
	External criterion (contrasting groups)
	Other

* Evidence marked with * refers to examples of analytical evidence.

Coding

Two researchers completed a data extraction form for every article. Ten percent of the articles were double-coded. An inter reliability analysis was subsequently performed to determine the consistency between the two raters, which was found to be high (Cohen's $\kappa = .89$, $p < .001$).

Analysis

Cross tabulations were made to compare the validation practices described in the articles. Chi-square statistics were also calculated to evaluate whether the sources of validity evidence reported in the articles differed for the different domains, educational levels, and most importantly, the intended use of the tests. For these analyses, the distribution of evidence over the inferences within a variable were compared with the total distribution of evidence. For example, the distribution of evidence reported in articles describing validations of language tests was compared with the distribution of evidence reported in all articles. The chi-square test was repeated for all domains, educational levels, and intended uses separately.

Results

In this section, we will first give a general overview of the selected articles. We will subsequently present the results in relation to the three research questions.

Description

A total of 178 articles were included in this study, the full list of which can be found in Appendix II. These articles described tests in different fields (domains). The domain containing the most articles was "Language" (Table 5.3). The domains "Mathematics" and "Science" were quite small in comparison with other domains since they only included 11 and eight articles, respectively. Therefore, the articles on tests in mathematics and science were grouped together with those classified within the domain "Other." The "Other" category, which includes mathematics and science, consisted of 41 articles.

An article could only be classified under a single domain. However, the test described in the article can be appointed to multiple educational levels and intended uses. In the sample of articles, 181 different educational levels are

mentioned. The tests described in the articles consisted of a total of 185 different intended uses.

Table 5.3 also shows the distribution of the 178 articles across the domains and the 185 named test usages. The table also shows that most articles were intended for selection purposes (mostly within the domains "Combination" and "Medical") and measures of progress (mostly "Language").

Table 5.3: Selected articles – distribution over domains and intended uses of assessment

	Language	Medical	Combination	Math	Science	Other	Total
Placement	7	2	2	2			13
Diagnosis	16	3	4	4			27
Selection	1	11	27			3	42
Classification	4	9	1		1	2	17
Progress	16	9	5	4		7	41
Other	5	2	3	1	2	2	15
Unknown	10	5	2		5	8	30
Total	59	41	44	11	8	22	

The articles describing tests for selection purposes were, for example, about the Scholastic Aptitude Test ("Combination") or, in a medical education context, admission to medical school. A similar trend can be observed in Table 5.4 where the majority of articles describe university level tests. This is consistent with the finding that most articles describe selection tests within the domains "Combination" or "Medical."

Table 5.4: Selected articles – distribution over domains and educational level

	Preschool	Primary school	Secondary school	University	Unknown	Total
Language	10	25	8	13	2	58
Medical	0	0	0	39	1	40
Combination	10	9	4	22	1	46
Math	2	6	1	2	0	11
Science	0	0	7	1	0	8
Other	0	3	5	10	3	21
Total	22	43	25	87	7	

Sources of validity evidence

This section describes the results corresponding to the first research question:

1. What specific sources of validity evidence do researchers report in articles on validation studies on educational assessments?

Table 5.5: Number of articles reporting particular sources of evidence

Inference	Evidence	Number of articles
1. Scoring	Scoring is correct	10
	*Rating scheme available	8
	Rater agreement	14
	Inter-rater reliability	30
	Intra-rater reliability	7
	Other	0
2. Generalization	Reliability coefficient	52
	Generalizability coefficient	18
	*Test blueprint	10
	Other	1
3. Extrapolation I	Construct underrepresentation	5
	Construct irrelevant variance	38
	External criterion (other test)	66
	Theoretical model of construct Factor analysis IRT analysis (e.g., calibration)	62
	*Authenticity	7
	*Cognitive labs	3
	Other	2
4. Extrapolation II	*Critical tasks	2
	*Stakeholders	22
	External criterion (prediction)	65
	Other	0
5. Decision	Norm groups	0
	*Standard setting	1
	Cut score	16
	External criterion (contrasting groups)	6
	Other	0
Total		445

* Evidence marked with * refers to examples of analytical evidence.

The data extraction form used to analyze the articles included several sources of evidence clustered under the five inferences. Table 5.5 displays the number of times a source of evidence was reported in an article.
One article may contain multiple sources of evidence, all of which are included in this table; therefore, the total of 445 sources of evidence outnumbers the total number of articles (178).
The evidence appearing most frequently in the included articles concerned the relationships between tests and an external criterion. The articles' authors presented these relationships to demonstrate convergent or discriminant validity in order to support extrapolation outside the specific test context (extrapolation I) or to show whether a test predicts future performance (extrapolation II). Another recurring source of evidence came from studies aimed at testing hypotheses on a theoretical construct (extrapolation I) – for example, by means of a factor analysis.
Validity evidence consists of both empirical and analytical evidence. Among the articles included in this study, 10% (43) of the reported evidence was considered analytical, and 90% (402) was empirical.

Sources of evidence in interpretive argument

This section discusses the results pertaining to the second research question:

2. Can we structure these sources under the inferences specified within the general form of an interpretive argument?

The sources of evidence listed in Table 5.2 were clustered according to the model of an interpretive argument with five inferences. To address the second research question, Table 5.6 shows the number of articles reporting one or more sources of evidence within an inference. We see that from the total of 178 articles, 52 reported at least one source of evidence for the first inference. In terms of inferences supported with evidence of tests in all domains, the third inference was most evidenced with 132 articles, and the fourth inference came in second with 85 articles.
If we look at articles describing the validation of tests that include a combination of constructs, we see an inverse picture: most articles presented evidence of the fourth inference, and the third inference came in second. The articles presenting validity evidence for tests from other domains did not differ significantly.

Table 5.6: Articles reporting at least one source of evidence in an inference

	Language	Medical	Combination**	Other	Total
1. Scoring	43% (23)	25% (10)	16% (7)	30% (12)	29% (52)
2. Generalization	48% (26)	40% (16)	16% (7)	50% (20)	39% (69)
3. Extrapolation I	85% (46)	58% (23)	59% (26)	93% (37)	74% (132)
4. Extrapolation II	30% (16)	53% (21)	73% (32)	40% (16)	48% (85)
5. Decision	19% (10)	15% (6)	5% (2)	5% (2)	11% (20)

** $x^2(4) \approx 18{,}02;\ p < 0{,}001$

For an interpretive argument, whether an inference is backed by evidence is not the only relevant consideration; the particular combination in which inferences are evidenced is also important. Figure 5.2 shows the number of articles reporting evidence for a certain combination of inferences. Only the combinations occurring at least once are displayed; there was no combination between generalization (inference 2) and extrapolation II (inference 4) in the considered articles, hence their absence from Figure 5.2.

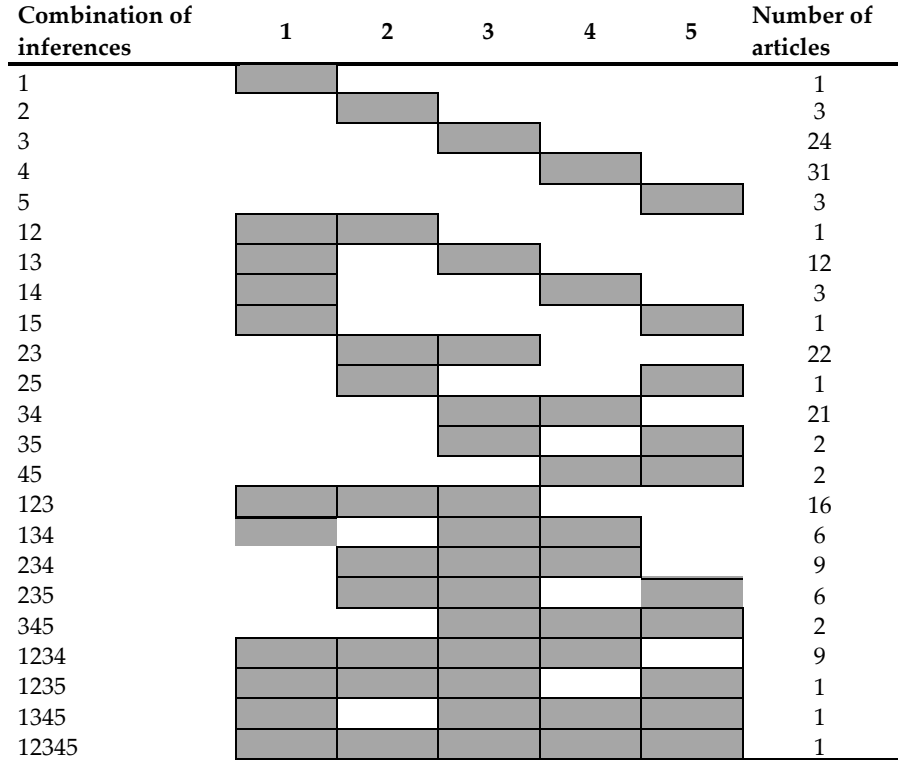


Figure 5.2: Combinations of inferences

Figure 5.2 shows, for example, that the majority of the articles reported evidence in extrapolation I (n=24), extrapolation II (n=31), or a combination of both inferences (n=22). It also shows that 127 articles reported evidence for only one or two inferences. Only one (Denton, Ciancio, & Fletcher, 2006) of the articles reported evidence for all inferences (1, 2, 3, 4, and 5).
Table 5.7 displays the proportion of articles reporting evidence for two or less inferences and for three or more inferences for the total sample as well as for all domains individually. These proportions do not differ significantly across domains.

Table 5.7: Percentage of articles reporting evidence within a certain number of inferences

	Language	Medical	Combination	Other	Total
1 or 2 inferences filled	63%	75%	86%	63%	71%
3 or more inferences filled	37%	25%	14%	38%	29%

Validity evidence for different intended uses of assessments

This section discusses the results of the third research question:

3. Are there differences between the sources of validity evidence gathered and reported for various kinds of interpretations and uses of tests?

In this section, we specifically look into the different intended uses of assessments in the selected articles. We hypothesized that the evidenced inferences would vary according to different intended uses. Table 5.8 displays the number of articles reporting evidence for an inference for the different intended uses discussed in the introduction of this article.

Table 5.8: Articles reporting evidence in an inference (per intended use of the assessment)

	Placement	**Diagnosis**	**Selection****	**Classification**	**Progress**
1. Scoring	2	11	3	5	14
2. Generalization	4	14	4	9	18
3. Extrapolation I	**9**	**21**	18	**14**	**34**
4. Extrapolation II	**7**	11	**36**	6	15
5. Decision	2	6	2	3	6
	13	27	42	17	41

** $(x^2(4) \approx 41{,}3;\ p < 0{,}001)$.

Table 5.8 lists the number of articles presenting validity evidence for a specific purpose. The shaded cells are hypothesized in Table 5.1 as most important for a specific use; the **bold** text indicates the inferences found to be most evidenced

within an intended use. We observed that articles describing validation efforts for selection tests differed significantly in the evidence presented from articles reporting validity evidence in other domains. For other purposes, the hypothesis was not confirmed since the articles presenting evidence for these purposes did not differ significantly from the total sample of articles.

As in research question two, the figures displaying the combination of evidenced inferences can be generated according to the intended use of the test (Figure 5.3). From this figure, it becomes apparent that articles for selection tests report evidence for only one or two inferences significantly more often than articles on other tests.

						Number of articles:				
Combination of inferences	**1**	**2**	**3**	**4**	**5**	**Placement**	**Diagnosis**	**Selection**	**Classification**	**Progress**
1	■							1		
2		■					1	1	1	
3			■			1	3	3	3	8
4				■		3	2	21	2	3
5					■	1	1			
12	■	■								
13	■		■				3			3
14	■			■				1		2
15	■				■					1
23		■	■			1	2		2	7
25		■			■		1			
34			■	■		4	2	11	1	1
35			■		■		1			1
45				■	■		1			1
123	■	■	■			2	3		3	4
134	■		■	■					1	2
234		■	■	■			1	2		3
235		■	■		■	1	1	1	2	2
345			■	■	■				1	1
1234	■	■	■	■			4		1	2
1235	■	■	■		■					
1345	■		■	■	■			1		
12345	■	■	■	■	■		1			

1. Scoring; 2. Generalization; 3. Extrapolation I; 4. Extrapolation II; 5. Decision.

Figure 5.3: Combinations of inferences and five intended uses

This result is also reflected in Table 5.9, which shows the percentage of articles presenting evidence for one or two inferences versus articles presenting evidence for three or more inferences. This table shows that selection tests differ significantly from tests with other purposes. The authors describing these tests tended to report only one or two sources of evidence. While this trend was also seen in relation to other tests, it was not as strong as for selection tests. Particularly, tests for classification purposes were often validated by means of evidence relating to more than two inferences.

Table 5.9: Percentage of articles reporting evidence within a certain number of inferences (per purpose)

	Placement	Diagnosis	Selection*	Classification	Progress
1 or 2 inferences filled	77%	63%	90%	53%	66%
3 or more inferences filled	23%	37%	10%	47%	34%

* $(x^2(4) \approx 38{,}0;\ p < 0{,}05)$.

Conclusion and discussion

The aim of this study was to create an inventory of current practices in validation studies. Three research questions were posed in relation to the argument-based approach to validation. The first question addressed the sources of validity evidence reported in articles on validation studies. Much of the reported evidence consisted of relationships with external criteria, such as other tests, expert judgments on competences, or the Grade Point Average (GPA). The final selection of articles pointed to different reasons for reporting on a relation with an external criterion: (1) to show a relation with a different construct or using the same construct to show that a test measures the correct construct; (2) to make a prediction of the future, for example, about the GPA; (3) to determine a cut score with a contrasting group's method (Hambleton & Pitoniak, 2006). Another very common source of evidence in the articles had to do with supporting the theoretical rationale regarding the construct. To do this, many authors performed a factor analysis to show that subscales can be distinguished within a construct. The third source of evidence, which was very often reported was a reliability coefficient. However, this coefficient was often not reported as part of validity evidence but as a separate measure of reliability. In his description of the argument-based approach to validation, Kane (2006) interprets reliability coefficients as a component of validity evidence, more

specifically, as part of the evidence that supports a generalization inference. In this study, we followed Kane and interpreted reliability as a possible source of evidence that can be presented in support of the validity of test scores. We therefore coded these coefficients as part of the reported evidence.

The second question was aimed at structuring the sources of evidence by using the inferences of an interpretive argument. The actual classification was done as part of the development of the data extraction form. The amount of evidence reported within the inferences and the combination of inferences evidenced in the articles were shown in the results section. When we examined the differences between assessments within the different test domains, we found that authors who presented evidence for assessments within the "Combination" domain reported significantly more evidence in the extrapolation II inference (competence domain – practice domain) while the others reported more evidence for the extrapolation I inference (test domain – competence domain).

In general, it was found that especially extrapolation I and II inferences received much attention. It is no coincidence that these inferences were precisely those including two of the three most reported sources of evidence (factor analysis and studies examining the relationship between a test and an external criterion). These particular inferences corresponded most with Messick's validity theory on construct validity (Messick, 1989) and commonly made up part of the evidence presented in relation to this conceptualization.

For the second research question, we also looked at the combination of evidence reported. The study did find one article that described a validity argument consisting of evidence from all inferences and emphasizing the most important ones (Denton, Ciancio, & Fletcher, 2006). However, in most of the included articles, only one or two inferences were evidenced. It seems that many articles tended to focus on a single element within the interpretive argument. It could be though that if all validation efforts pertaining to a particular test were combined, we would see a different pattern, for example, that once combined, all inferences would receive attention. If this was the case, these articles did not appear in our sample since none of the articles summarized all available evidence into one coherent validity argument.

The third question addressed the intended uses of assessments. Remarkably, we found that 30 out of the 178 articles did not specifically report the intended use of the validated test. However, upon examination of those articles that did provide information on intended use, we observed that the reported evidence did differ for different uses. These differences were most prominent for tests with selection purposes, and the validity evidence presented to support these

tests did differ significantly from the evidence presented for tests with other purposes. This is in congruence with the earlier finding that tests within the "Combination" domain differ from those in other domains. Many tests with selection purposes (e.g., admission tests) cover multiple constructs and are therefore classified within the "Combination" domain. In general, we expected some specific inferences to be emphasized for specific purposes, but they were not always evidenced. This might have been because we found that regardless of the intended use, the third and fourth inferences received the most attention.

Discussion

In this study, we found that the argument-based approach to validation was not a very common conceptualization of validation studies in recently published articles. It appears that researchers and practitioners continue to present evidence under the conceptualization of validity proposed by Messick (1989). Unfortunately, this also means that the presented evidence is not always the most relevant for validating a specific interpretation or use of test scores.

Notwithstanding, in relation to the current study, we maintain that the conclusions formulated above are somewhat biased because of the inclusion criteria used. Due to the decision to limit the search to peer-reviewed journals, many validation studies reported in research reports or theses were left out. Furthermore, it appears uncommon to publish every aspect of test development in peer-reviewed journals. This could be due to publication bias in relation to new methods and tests as well as the fact that existing validation practice does not meet these criteria. Furthermore, it might be the case that validity evidence is collected by test publishers but that they do not feel the need to publish this evidence as part of the scientific discourse.

Another limitation of this study is the classification of evidence within the specific elements of the interpretive argument defined in the argument-based approach. Many authors were not very specific about why they presented their sources of evidence; therefore, it was difficult to identify the inferences relating to the evidence. This means that sometimes, in the classification of evidence, assumptions were made regarding the rationale of collecting specific sources of validity evidence. Therefore, it is possible to debate some of the choices made in this classification. In this study, evidence might, for example, be regarded as evidence of construct validity while, in reality, researchers might have conducted a study to support a decision-making claim. This does not mean that the conclusions drawn in this article are invalid. The results seem to be clear in

the sense that the patterns found in this study would most likely also occur if other choices were made in the classification of evidence.

In conclusion, in this study, we showed that building arguments and specifically evidencing weaker inferences have not yet emerged as daily practice for validation studies. Authors still seemed to rely on validation theory, as proposed by Messick (1989), and less on new insights in argument-based theories (Kane, 2013). This is unfortunate since the argument-based approach offers researchers and test developers a framework to guide validation research in the direction where it is most needed: the weakest inferences in their validity claim. It could be that researchers and test developers are still unfamiliar with this method, in which case, it might help to publish examples of validation studies that follow the argument-based approach to validation (Wools, Béguin & Eggen, submitted) or to develop tools that support validation practice by means of an argument-based approach (Wools, 2012).

References

American Educational Research Association (AERA). American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999), *Standards for educational and psychological testing*. Washington: American Psychological Association.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1–34.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly,* 41(1), 8–34.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (1992) An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2004). Certification testing as an Illustration of argument-based validation, *Measurement*, *2*, 135–70.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). Charlotte, NC: Information Age Pub.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Lissitz, R. W. (2009). *The concept of validity*. Charlotte: IAP-Information Age Publishing.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, *27*(3), 32–42.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students.* Upper Saddle River, NJ: Pearson/Merrill Prentice Hill.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden: Blackwell Publishing.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Washington DC: American Council on Education.

Wools, S., Béguin, A., Eggen, T. (submitted). Comparable performance standards in arithmetic for different educational tracks in the Netherlands.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63–82.

Wools, S. (2012). Towards a comprehensive evaluation system for the quality of tests and assessments. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 95–106). Enschede: RCEC.

Appendix I

Data extraction form

General:

Title of the study	
Author(s)	
Journal, Volume, pages	
Year of publication	
The research goal as described by the authors/research question	

About validated test:

Domain	Language Medical Combination Math Science Other
Construct	
Education level or stream (+ grade level)	Preschool Primary school Secondary school University Unknown
Purpose of the test	Placement Diagnosis Selection Classification Progress Other Unknown
Conclusion about validity	Yes In between No Unknown

Additional comments:

Validity Evidence:

Inference	Evidence	Yes?
1. performance - score	Scoring is correct	
	Rating scheme available	
	Rater agreement	
	Inter-rater reliability	
	Intra-rater reliability	
	Other	
2. score – test domain	Reliability coefficient	
	Generalizability coefficient	
	Test blueprint	
	Other	
3. test domain – competence domain	Construct underrepresentation	
	Construct irrelevant variance	
	External criterion (other test)	
	Theoretical model of construct Factor analysis IRT analysis (e.g., calibration)	
	Authenticity	
	Cognitive labs	
	Other	
4. competence domain – practice domain	Critical tasks	
	Stakeholders	
	External criterion (prediction)	
	Other	
5. practice domain – decision	Norm groups	
	Standard setting	
	Cut score (right score)	
	External criterion (contrasting groups)	
	Other	

Appendix II

Articles included in the systematic literature review (178).

Admiraal, W., Hoeksma, M., van de Kamp, M.-T., & van Duin, G. (2011). Assessment of teacher competence using video portfolios: Reliability, construct validity, and consequential validity. *Teaching and Teacher Education*, *27*(6), 1019–1028.

Albers, C. A. (2012). Alternate English language proficiency assessment for ELLs with significant disabilities: Validity evidence from alignment with English language proficiency standards. *The International Journal of Educational and Psychological Assessment*, *10*(1), 97–124.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*(1), 31–47.

Ansari, A. A., Ali, S. K., & Donnon, T. (2013). The construct and criterion validity of the mini-CEX: A meta-analysis of the published research. *Academic Medicine*, *88*(3), 413–420.

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement*, *69*(6), 978–993.

Augustyniak, K. M., Cook-Cottone, C. P., & Calabrese, N. (2004). The predictive validity of the Phelps Kindergarten Readiness Scale. *Psychology in the Schools*, *41*(5), 509–516.

Baker, S. K., & Good, R. H. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review*, *24*(4), 561–578.

Ballantine, J. A., Larres, P. M., & Oyelere, P. (2007). Computer usage and the validity of self-assessed computer competence among first-year business students. *Computers & Education*, *49*(4), 976–990.

Balogh, J., Bernstein, J., Cheng, J., Van Moere, A., Townshend, B., & Suzuki, M. (2012). Validation of automated scoring of oral reading. *Educational and Psychological Measurement*, *72*(3), 435–452.

Banerji, M. (1999). Validation of scores/measures from a K-2 developmental assessment in mathematics. *Educational and Psychological Measurement*, *59*(4), 694–715.

Banerji, M., & Ferron, J. (1998). Construct validity of scores on a developmental assessment with mathematical patterns tasks. *Educational and Psychological Measurement*, *58*(4), 634–660.

Bardes, C. L., Colliver, J. A., Alonso, D. R., & Swartz, M. H. (1996). Validity of standardized-patient examinations scores as an indicator of faculty observer ratings. *Academic Medicine*, *71*(Suppl 1), S82–S83.

Barghaus, K. M., & Fantuzzo, J. W. (2014). Validation of the Preschool Child Observation Record: Does it pass the test for use in Head Start? *Early Education and Development*, 25(8), 1118–1141.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118.

Bennett, R. E., Gottesman, R. L., Cerullo, F. M., & Rock, D. A. (1991). The validity of Einstein Assessment subtest scores as predictors of early school achievement. *Journal of Psychoeducational Assessment*, *9*(1), 67–79.

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test. *Journal of Educational Measurement*, *32*(1), 19–36.

Berent, G. P., Samar, V. J., Kelly, R. R., Berent, R., Bochner, J., Albertini, J., & Sacken, J. (1996). Validity of indirect assessment of writing competency for deaf and hard-of-hearing college students. *Journal of Deaf Studies and Deaf Education*, *1*(3), 167–178.

Betts, J., Pickart, M., & Heistad, D. (2009). Construct and predictive validity evidence for curriculum-based measures of early literacy and numeracy skills in kindergarten. *Journal of Psychoeducational Assessment*, *27*(2), 83–95.

Bezruczko, N. (1995). Validation of a multiple choice visual arts achievement test. *Educational and Psychological Measurement*, *55*(4), 664–674.

Bogo, M., Regehr, C., Hughes, J., Power, R., & Globerman, J. (2002). Evaluating a measure of student field performance in direct service: Testing reliability and validity of explicit criteria. *Journal of Social Work Education*, *38*(3), 385–401.

Bouter, S., van Weel-Baumgarten, E., & Bolhuis, S. (2013). Construction and validation of the Nijmegen Evaluation of the Simulated Patient (NESP): Assessing simulated patients' ability to role-play and provide feedback to students. *Academic Medicine*, *88*(2), 253–259.

Brown, C. R., & Moore, J. L. (1994). Construct validity and context dependency of the assessment of practical skills in an advanced level biology examination. *Research in Science & Technological Education*, *12*(1), 53–61.

Brown, C. R., Pacini, D. J., & Taylor, D. J. (1992). Two different methods of assessing practical skills at an advanced level examination in biology: Demonstration of construct validity or the appraisal of non-events? *Research in Science & Technological Education*, *10*(1), 23–35.

Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning*, *42*(3), 313–357.

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (2013). Measuring students' writing ability on a Computer-Analytic Developmental Scale: An exploratory validity study. *Literacy Research and Instruction*, *52*(4), 255–280.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice*, *13*(1), 9–15.

Cannon, J. E., & Hubley, A. M. (2014). Content validation of the Comprehension of Written Grammar Assessment for deaf and hard of hearing students. *Journal of Psychoeducational Assessment*, *32*(8), 768–774.

Chang, C.-C., Tseng, K.-H., Chou, P.-N., & Chen, Y.-H. (2011). Reliability and validity of web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education*, *57*(1), 1306–1316.

Chen, Y.-H., Gorin, J. S., Thompson, M. S., & Tatsuoka, K. K. (2008). Cross-cultural validity of the TIMSS-1999 Mathematics Test: Verification of a cognitive model. *International Journal of Testing*, *8*(3), 251–271.

Chernyshenko, O. S., & Ones, D. S. (1999). How selective are psychology graduate programs? The effects of the selection ratio on GRE score validity. *Educational and Psychological Measurement*, *59*(6), 951–961.

Christ, T. J., White, M. J., Ardoin, S. P., & Eckert, T. L. (2013). Curriculum based measurement of reading: Consistency and validity across best, fastest, and question reading conditions. *School Psychology Review*, *42*(4), 415–436.

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment*, *13*(1), 56–75.

Collins, R., Elliot, N., Klobucar, A., & Deek, F. P. (2012). Web-based portfolio assessment: Validation of an open source platform. *Journal of Interactive Learning Research*, *24*(1), 5–32.

Crawford, L., Tindal, G., & Carpenter, D. M., II. (2006). Exploring the validity of the Oregon Extended Writing Assessment. *The Journal of Special Education*, *40*(1), 16–27.

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational and Psychological Measurement*, *61*(5), 841–848.

Crisp, V., & Shaw, S. (2012). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies*, *38*(2), 209–222.

Daly, E. J., III, Wright, J. A., Kelly, S. Q., & Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly*, *12*(3), 268–280.

Daub, D., & Colarusso, R. P. (1996). The validity of the WJ—R, PIAT—R, and DAB—2 reading subtests with students with learning disabilities. *Learning Disabilities Research & Practice*, *11*(2), 90–95.

Davies, P. (2004). Don't write, just mark: The validity of assessing student ability via their computerized peer-marking of an essay rather than their creation of an essay. *ALT-J Association for Learning Technology Journal*, *12*(3), 261–277.

de Lima, A. A., Barrero, C., Baratta, S., Costa, Y. C., Bortman, G., Carabajales, J., . . . Van der Vleuten, G. (2007). Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Medical Teacher*, *29*(8), 785–790.

Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly*, *41*(1), 8–34.

Diamantopoulou, S., Pina, V., Valero-Garcia, A. V., González-Salinas, C., & Fuentes, L. J. (2012). Validation of the Spanish version of the Woodcock-Johnson Mathematics Achievement Tests for children aged 6 to 13. *Journal of Psychoeducational Assessment*, *30*(5), 466–477.

Dobson, P., Krapljan-Barr, P., & Vielba, C. (1999). An evaluation of the validity and fairness of the Graduate Management Admissions Test (GMAT) used for MBA selection in a UK business school. *International Journal of Selection and Assessment*, *7*(4), 196–202.

Dong, T., Swygert, K. A., Durning, S. J., Saguil, A., Gilliland, W. R., Cruess, D., . . . Artino, A. R., Jr. (2014). Validity evidence for medical school OSCEs: Associations with USMLE® step assessments. *Teaching and Learning in Medicine*, *26*(4), 379–386.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true–false items in medical specialty certification. *Applied Measurement in Education*, *8*(2), 187–197.

Driessen, E. W., Overeem, K., van Tartwijk, J., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2006). Validity of portfolio assessment: Which qualities determine ratings? *Medical Education*, *40*(9), 862–866.

Duncan, J., & Rafter, E. M. (2005). Concurrent and predictive validity of the Phelps Kindergarten Readiness Scale-II. *Psychology in the Schools*, *42*(4), 355–359.

Dunleavy, D. M., Kroopnick, M. H., Dowd, K. W., Searcy, C. A., & Zhao, X. (2013). The predictive validity of the MCAT exam in relation to academic performance through medical school: A national cohort study of 2001–2004 matriculants. *Academic Medicine*, *88*(5), 666–671.

Durning, S. J., Artino, A., Boulet, J., La Rochelle, J., Van Der Vleuten, C., Arze, B., & Schuwirth, L. (2012). The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. *Medical Teacher*, *34*(1), 30–37.

Edgcumbe, D. P., Silverman, J., & Benson, J. (2012). An examination of the validity of EPSCALE using factor analysis. *Patient Education and Counseling*, *87*(1), 120–124.

Edwards, W. R., & Schleicher, D. J. (2004). On selecting psychology graduate students: Validity evidence for a Test of Tacit Knowledge. *Journal of Educational Psychology*, *96*(3), 592–602.

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills—Modified. *School Psychology Review*, *30*(1), 33–49.

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups, multimethod approach. *Educational Measurement: Issues and Practice*, *26*(2), 30–43.

Emery, J. L., & Bell, J. F. (2009). The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education*, *43*(6), 557–564.

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, 16(3), 142-151.

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *15*(1), 5–27.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, *10*(3), 274–291.

Farr, R., & Jongsma, E. (1993). The convergent/discriminant validity of integrated reading/writing assessment. *Journal of Research & Development in Education*, *26*(2), 83–91.

Fewster, S., & MacMillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education*, *23*(3), 149–156.

Frisbie, D. A., & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement*, *32*(1), 55–78.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended computerized mathematics task. *Educational Assessment*, *8*(1), 27–41.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, *31*(4), 477–497.

Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., & Naquin, G. M. (2004). An examination of the Criterion Validity and Sensitivity to Brief Intervention of Alternate Curriculum-Based Measures of Writing Skill. *Psychology in the Schools*, *41*(3), 291–300.

García-Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university context. *Electronic Journal of Research in Educational Psychology*, *9*(3), 1043–1062.

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8*(1), 1–26.

Goffreda, C. T., Diperna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools*, *46*(6), 539–552.

Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement*, *52*(4), 1019–1027.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment*, *12*(3-4), 239–266.

Guerrero, M. D. (2000). The unified validity of the Four Skills Exam: Applying Messick's framework. *Language Testing*, *17*(4), 397–421.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*(3), 423–449.

Hager, K. D., & Slocum, T. A. (2008). Utah's Alternate Assessment: Evidence regarding six aspects of validity. *Education and Training in Developmental Disabilities*, *43*(2), 144–161.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: II. NELS: 88 science achievement. *American Educational Research Journal*, *32*(3), 555–581.

Havey, J. M., Story, N., & Buker, K. (2002). Convergent and concurrent validity of two measures of phonological processing. *Psychology in the Schools*, *39*(5), 507–514.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*(3), 201–224.

Hewitt, M. A., & Homan, S. P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction*, *43*(2), 1–16.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review*, *32*(4), 541–556.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a Story Retelling Speaking Test. *Language Assessment Quarterly*, *10*(4), 398–422.

Hodges, B., Regehr, G., Hanson, M., & McNaughton, N. (1998). Validation of an objective structured clinical examination in psychiatry. *Academic Medicine*, *73*(8), 910–912.

House, J. D., & Johnson, J. J. (2002). Predictive validity of the Graduate Record Examination Advanced Psychology Test for grade performance in graduate psychology courses. *College Student Journal*, *36*(1), 32–36.

Huff, K. L., Koenig, J. A., Treptau, M. M., & Sireci, S. G. (1999). Validity of MCAT scores for predicting clerkship performance of medical students grouped by sex and ethnicity. *Academic Medicine*, *74*(Suppl 10), S41–S44.

Huggins, A. C., Ritzhaupt, A. D., & Dawson, K. (2014). Measuring information and communication technology literacy using a performance assessment: Validation of the student tool for technology literacy (ST²L). *Computers & Education*, *77*, 1–12.

Husbands, A., & Dowell, J. (2013). Predictive validity of the Dundee Multiple Mini-Interview. *Medical Education*, *47*(7), 717–725.

Jansen, J. J. M., Scherpbier, A. J. J. A., Metz, J. C. M., Grol, R. P. T. M., & van der Vleuten, C. P. M. (1996). Performance-based assessment in continuing medical education for general practitioners: Construct validity. *Medical Education*, *30*(5), 339–344.

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, *34*(3), 358–371.

Kaminski, R. A., Abbott, M., Bravo Aguayo, K., Latimer, R., & Good, R. H., III. (2014). The Preschool Early Literacy Indicators: Validity and benchmark goals. *Topics in Early Childhood Special Education*, *34*(2), 71–82.

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning*, *25*(3), 257–275.

Kobrin, J. L., Kim, Y., & Sackett, P. R. (2012). Modeling the predictive validity of SAT mathematics items using item characteristics. *Educational and Psychological Measurement*, *72*(1), 99–119.

Kobrin, J. L., & Patterson, B. F. (2011). Contextual factors associated with the validity of SAT scores and high school GPA for predicting first-year college grades. *Educational Assessment*, *16*(4), 207–226.

Koenig, J. A., Sireci, S. G., & Wiley, A. (1998). Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine*, *73*(10), 1095–1106.

Kramer, A. W. M., Jansen, J. J. M., Zuithoff, P., Düsman, H., Tan, L. H. C., Grol, R. P. T. M., & van der Vleuten, C. P. M. (2002). Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. *Medical Education*, *36*(9), 812–819.

Kreiter, C. D., & Kreiter, Y. (2007). A validity generalization perspective on the ability of undergraduate GPA and the Medical College Admission Test to predict important outcomes. *Teaching and Learning in Medicine*, *19*(2), 95–100.

Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist*, *53*(5), 567–568.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education*, *6*(1), 51–68.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, *70*(2), 340–352.

Kyriakides, L. (2004). Investigating validity from teachers' perspectives through their engagement in large-scale assessment: The Emergent Literacy Baseline Assessment Project. *Assessment in Education: Principles, Policy & Practice*, *11*(2), 143–165.

Lane, S., Lui, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a Mathematics Performance Assessment. *Journal of Educational Measurement*, *33*(1), 71–92.

Lawlor, S., Richman, S., & Richman, C. L. (1997). The validity of using the SAT as a criterion for black and white students' admission to college. *College Student Journal*, *31*(4), 507–515.

Lee, H.-K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), 307–330.

Lee Webb, M.-Y., Schwanenflugel, P. J., & Kim, S.-H. (2004). A construct validation study of phonological awareness for children entering prekindergarten. *Journal of Psychoeducational Assessment*, *22*(4), 304–319.

Leighton, J. P., Heffernan, C., Cor, M. K., Gokiert, R. J., & Cui, Y. (2011). An experimental test of student verbal reports and teacher evaluations as a source of validity evidence for test development. *Applied Measurement in Education*, *24*(4), 324–348.

Li, Y. H., & Tompkins, L. J. (2004). Examining the construct validity for the Multiple-Content Testing Programs. *International Journal of Testing*, *4*(3), 217–238.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*(4), 245–257.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, *47*(12), 1175–1183.

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, *48*(9), 1079–1107.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, *27*(3), 32–42.

Lockspeiser, T. M., Schmitter, P. A., Lane, J. L., Hanson, J. L., Rosenberg, A. A., & Park, Y. S. (2013). Assessing residents' written learning goals and goal writing skill: Validity evidence for the Learning Goal Scoring Rubric. *Academic Medicine*, *88*(10), 1558–1563.

Lohfeld, L., Goldie, J., Schwartz, L., Eva, K., Cotton, P., Morrison, J., . . . Wood, T. (2012). Testing the validity of a scenario-based questionnaire to assess the ethical sensitivity of undergraduate medical students. *Medical Teacher*, *34*(8), 635–642.

Madeline, A., & Wheldall, K. (1998). Towards a curriculum-based passage reading test for monitoring the performance of low-progress readers using standardized passages: A validity study. *Educational Psychology*, *18*(4), 471–478.

Martin, I. G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. *Medical Education*, *36*(5), 418–425.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice*, *23*(4), 16–30.

Massey, A. J. (1997). Multitrait-multimethod/multiform evidence for the validity of reporting units in national assessments in science at age 14 in England and Wales. *Educational and Psychological Measurement*, *57*(1), 108–117.

Matsell, D. G., Wolfish, N. M., & Hsu, E. (1991). Reliability and validity of the objective structured clinical examination in paediatrics. *Medical Education*, *25*(4), 293–299.

Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2011). An alternative presentation of incremental validity: Discrepant SAT and HSGPA performance. *Educational and Psychological Measurement*, *71*(4), 638–662.

McMurray, M. A., Beisenherz, P., & Thompson, B. (1991). Reliability and concurrent validity of a measure of critical thinking skills in biology. *Journal of Research in Science Teaching*, *28*(2), 183–191.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to Grade 3. *American Educational Research Journal*, *38*(1), 73–95.

Mercer, S. H., Martínez, R. S., Faust, D., & Mitchell, R. R. (2012). Criterion-related validity of curriculum-based measurement in writing with narrative and expository prompts relative to passage copying speed in 10th grade students. *School Psychology Quarterly*, *27*(2), 85–95.

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review*, *37*(3), 359–373.

Meyer, J. H., Woodard, P. G., & Suddick, D. E. (1994). The Descriptive Tests of Mathematics Skills: Predictive validity for an elementary mathematics concepts and structures course. *Educational and Psychological Measurement*, *54*(1), 115–117.

Millett, J., Atwill, K., Blanchard, J., & Gorin, J. (2008). The validity of receptive and expressive vocabulary measures with Spanish-speaking kindergartners learning English. *Reading Psychology*, *29*(6), 534–551.

Mitchell, K., Haynes, R., & Koenig, J. (1994). Assessing the validity of the updated Medical College Admission Test. *Academic Medicine*, *69*(5), 394–401.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement*, *55*(2), 309–316.

Olsen, J. B., Cox, A., Price, C., Strozeski, M., & Vela, I. (1990). Development, implementation, and validation of a computerized test for statewide assessment. *Educational Measurement: Issues and Practice*, *9*(2), 7–10.

Olson, L. G. (2000). The effect of a Structured Question Grid on the validity and perceived fairness of a medical long case assessment. *Medical Education*, *34*(1), 46–52.

Oreck, B. A., Owen, S. V., & Baum, S. M. (2003). Validity, reliability, and equity issues in an observational talent assessment process in the performing arts. *Journal for the Education of the Gifted*, *27*(1), 62–94.

Panter, J. E. (2000). Validity of the Bracken Basic Concept Scale-Revised for predicting performance on the Metropolitan Rediness Test-Sixth Edition. *Journal of Psychoeducational Assessment*, *18*(2), 104–110.

Panter, J. E., & Bracken, B. A. (2009). Validity of the Bracken School Readiness Assessment for predicting first grade readiness. *Psychology in the Schools*, *46*(5), 397–409.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish Listening Exam: Test usefulness evaluation. *Language Assessment Quarterly*, *7*(2), 137–159.

Park, Y. S., Lineberry, M., Hyderi, A., Bordage, G., Riddle, J., & Yudkowsky, R. (2013). Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. *Academic Medicine*, *88*(10), 1552–1557.

Park, Y. S., Riddle, J., & Tekian, A. (2014). Validity evidence of resident competency ratings and the identification of problem residents. *Medical Education*, *48*(6), 614–622.

Petek, J. M., & Todd, W. F. (1991). Predictive validity of the new MCAT relative to other preadmission predictors in podiatric schools. *Academic Medicine*, *66*(7), 425–425.

Pomplun, M. (2004). The differential predictive validity of the Initial Skills Analysis: Reading screening tests for K-3. *Educational and Psychological Measurement*, *64*(5), 813–827.

Qi, C. H., & Marley, S. C. (2011). Validity study of the Preschool Language Scale-4 with English-Speaking Hispanic and European American children in head start programs. *Topics in Early Childhood Special Education*, *31*(2), 89–98.

Ravesloot, C., van der Schaaf, M., Haaring, C., Kruitwagen, C., Beek, E., Ten Cate, O., & van Schaik, J. (2012). Construct validation of progress testing to measure knowledge and visual skills in radiology. *Medical Teacher*, *34*(12), 1047–1055.

Ritchey, K. D., & Coker, D. L., Jr. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *29*(1), 89–119.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *The Journal of Special Education*, *38*(4), 218–231.

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential validity and utility of successive and simultaneous approaches to the development of Equivalent Achievement Tests in French and English. *Alberta Journal of Educational Research*, *49*(3), 290–304.

Rogers, W. T., Lin, J., & Rinaldi, C. M. (2011). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education*, *24*(1), 39–70.

Rothman, A. I., Cohen, R., Dirks, F. R., Poldre, P., & Ross, J. (1991). Validity and reliability of a domain-referenced test of clinical competence for foreign medical graduates. *Academic Medicine*, *66*(7), 423–425.

Rouse, H. L., & Fantuzzo, J. W. (2006). Validity of the Dynamic Indicators for Basic Early Literacy Skills as an indicator of early literacy for urban kindergarten children. *School Psychology Review*, *35*(3), 341–355.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355–390.

Schmidt, A. E. (2000). An approximation of a hierarchical logistic regression model used to establish the predictive validity of scores on a nursing licensure exam. *Educational and Psychological Measurement*, *60*(3), 463–478.

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: Development and validation. *Medical Teacher*, *30*(5), 528–533.

Segers, M., Dierick, S., & Dochy, F. (2001). Quality standards for new modes of assessment. An exploratory study of the consequential validity of the OverAll Test. *European Journal of Psychology of Education*, *16*(4), 569–588.

Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & McDougal, J. (2014). Concurrent and predictive validity of reading retell as a brief measure of reading comprehension for narrative text. *Reading Psychology*, *35*(7), 644–665.

Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching*, *34*(7), 721–743.

Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly*, *10*(4), 452–475.

Shen, W., Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., & Kiger, T. B. (2012). All validities are not created equal: Determinants of variation in SAT validity across schools. *Applied Measurement in Education*, *25*(3), 197–219.

Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement*, *66*(2), 305–317.

Strand, S. (2006). Comparing the predictive validity of reasoning tests and national end of Key Stage 2 tests: Which tests are the "best"? *British Educational Research Journal*, *32*(2), 209–225.

Streyffeler, L., Altmaier, E. M., Kuperman, S., & Patrick, L. E. (2005). Development of a medical school admissions interview phase 2: Predictive validity of cognitive and non-cognitive attributes. *Medical Education Online*, *10*, 1–5.

Stricker, L. J. (1991). Current validity of 1975 and 1985 SATs: Implications for validity trends since the mid-1970s. *Journal of Educational Measurement*, *28*(2), 93–98.

Talento-Miller, E., & Rudner, L. M. (2008). The validity of Graduate Management Admission Test scores: A summary of studies conducted from 1997 to 2004. *Educational and Psychological Measurement*, *68*(1), 129–138.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, *31*(4), 498–513.

Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research & Practice*, *24*(3), 132–142.

Tokar, E., & DeBlois, C. S. (1991). Evidence of the validity of the PPST for non-traditional college students. *Educational and Psychological Measurement*, *51*(1), 161–166.

Travis, T. A., Colliver, J. A., Robbs, R. S., Barnhart, A. J., Barrows, H. S., Giannone, L., . . . Steward, D. E. (1996). Validity of a simple approach to scoring and standard setting for standardized-patient cases in an examination of clinical competence. *Academic Medicine*, *71*(Suppl 1), S84–S86.

Tromp, F., Vernooij-Dassen, M., Grol, R., Kramer, A., & Bottema, B. (2012). Assessment of CanMEDS roles in postgraduate training: The validation of the Compass. *Patient Education and Counseling*, *89*(1), 199–204.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45*(3), 270–291.

Vallevand, A., & Violato, C. (2012). A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of international medical graduates. *Teaching and Learning in Medicine*, *24*(2), 168–176.

van Diepen, M., Verhoeven, L., Aarnoutse, C., & Bosman, A. M. T. (2007). Validation of the international Reading Literacy Test: Evidence from Dutch. *Written Language and Literacy*, *10*(1), 1–23.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411–440.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review*, *30*(3), 363–382.

Varkey, P., Natt, N., Lesnick, T., Downing, S., & Yudkowsky, R. (2008). Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: A preliminary investigation. *Academic Medicine*, *83*(8), 775–780.

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952–969.

Wainer, H., Saka, T., & Donoghue, J. R. (1993). The validity of the SAT at the University of Hawaii: A riddle wrapped in an enigma. *Educational Evaluation and Policy Analysis*, *15*(1), 91–98.

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review*, *62*(3), 323–336.

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*(3), 311–333.

Wijnen-Meijer, M., Van der Schaaf, M., Booij, E., Harendza, S., Boscardin, C., Van Wijngaarden, J., & Ten Cate, T. J. (2013). An argument-based approach to the validation of UHTRUST: Can we measure how recent graduates can be trusted with unfamiliar tasks? *Advances in Health Sciences Education*, *18*(5), 1009–1027.

Williams, V. S., & Wakeford, M. E. (1993). The predictive validity of the National Teacher Examinations used for admission to teacher education programs. *Educational and Psychological Measurement*, *53*(2), 533–539.

Willoughby, T. L., & Bixby, A. R. (1991). Cross-validation of the Quarterly Profile Examination. *Educational and Psychological Measurement*, *51*(3), 691–697.

Wilson, J., & Wright, C. R. (1993). The predictive validity of student self-evaluations, teachers' assessments, and grades for performance on the Verbal Reasoning and Numerical Ability scales of the Differential Aptitude Test for a sample of secondary school students attending rural Appalachia schools. *Educational and Psychological Measurement*, *53*(1), 259–270.

Yao, Y., Thomas, M., Nickens, N., Downing, J. A., Burkett, R. S., & Lamson, S. (2008). Validity evidence of an electronic portfolio for preservice teachers. *Educational Measurement: Issues and Practice*, *27*(1), 10–24.

Yarroch, W. L. (1991). The implications of content versus item validity on science tests. *Journal of Research in Science Teaching*, *28*(7), 619–629.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, *13*(2-3), 170–192.

Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation and Accountability*, *24*(3), 263–277.

Zwick, R. (1993). The validity of the GMAT for the prediction of grades in doctoral study in business and management: An empirical Bayes approach. *Journal of Educational Statistics*, *18*(1), 91–107.

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48*(2), 101–121.

Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice*, *23*(1), 6–16.

Chapter 6

Towards a comprehensive evaluation system for the quality of tests and assessments

Abstract

To evaluate the quality of educational assessments, several evaluation systems are available. These systems are, however, focused around the evaluation of a single type of test. Furthermore, within these systems, quality is defined as a non-flexible construct, whereas in this paper it is argued that the evaluation of test quality should depend on the test's purpose. Within this paper, we compare several available evaluation systems. From this comparison, design principles are derived to guide the development of a new, comprehensive quality evaluation system. The paper concludes with an outline of the new evaluation system, which intends to incorporate an argument-based approach to quality.

Standards, evaluation, quality, educational assessment, argument-based approach

Chapter previously published as:
Wools, S. (2012). Towards a Comprehensive Evaluation System for the Quality of Tests and Assessments. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC*. (pp. 95-106). Enschede: RCEC.

Introduction

In all levels of education, students have to take tests and assessments to demonstrate their ability, for example, to show whether they have fulfilled the course objectives or to guide them in their further learning. In the context of high-stakes exams and assessments, the importance of good quality decisions is clear. However, in other contexts, the assessment results need to be valid and reliable too. In other words, despite the stakes of an exam, the results need to be appropriate for its intended use. This can only occur when the assessment instruments that are used to assess the students are of good quality. To evaluate test quality, several evaluation systems and standards are available. The currently available evaluation systems, however, tend to focus around one specific type of test or test use, for example, computer-based tests (Keuning, 2004), competence-based assessments (Wools, Sanders, & Roelofs, 2007), examinations (Sanders, 2011), or psychological tests (Evers, Lucassen, Meijer, & Sijtsma, 2010). Standards are often more broadly defined, but are aimed at guiding test developers during the development process and are not suited for an external evaluation of quality.

The purpose of this paper is to introduce the outline of a new evaluation system that will be more flexible and comprehensive than the currently available evaluation systems. Furthermore, this proposed evaluation system is not only suitable to guide test development, but can also be used as an instrument for internal or external audits. In the first section of this paper, the available standards and evaluation systems are described. In the second section, the principles that serve as a basis for the new evaluation system are specified. From this second section, we will derive the design of the new system that is described in the final section of this paper.

Section 1 - Guidelines, standards, and evaluation systems

To describe the currently available systems for the evaluation of test quality, we will compare nine quality evaluation systems. The nine systems will be compared based on their purpose, their intended audience, and their object of evaluation. We do not aim to include all of the available evaluation systems, nor will we describe every aspect for every system that is mentioned, since this section is meant mainly to exemplify the diversity of the systems.

We will differentiate between guidelines, standards, and evaluation systems. Guidelines suggest quality aspects that you *can* comply with. Standards mention aspects of quality that you *should* comply with, in order to develop sound and reliable tests. Evaluation systems focus on evaluating a test, and prescribe what quality aspect *must* be met to ensure minimal quality. We will also add criteria to the comparison that are mentioned by researchers as being important, but that are not implemented in the guidelines, standards, or evaluation systems.

Systems for comparison

Guidelines:

1. International guidelines for test use from International Testing Committee (ITC) (Bartram, 2001)

Standards:

2. Standards for educational and psychological testing (AERA, APA, & NCME, 1999)
3. European framework of standards for educational assessment (AEA-Europe, 2012)
4. ETS standards (Educational Testing Service (ETS), 2002)
5. Cambridge approach (Cambridge Assessment, 2009)
6. Code of fair testing practices in education (Joint Committee on Testing Practices (JCTP), 2004).

Evaluation systems:

7. COTAN evaluation system for test quality (Evers et al., 2010)
8. EFPA review model for the description and evaluation of psychological tests (Lindley, Bartram, & Kennedy, 2004)

Criteria:

9. Quality criteria for competence assessment programs (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006)

Table 6.1: Comparison of standards, guidelines, and evaluation systems

	9. Baartman	8. EFPA	7. COTAN	6. JCTP	5.Cambridge Assessment	4. ETS	3.AEA-Europe	2.AERA Standards	1. ITC
Guide test development				x	x	x	x		
Guide test use				x			x	x	x
Guide self-evaluation	x						x		
Guide audits		x	x				x	x	
Test specialists		x	x				x	x	
Teachers	x			x			x		x
Users	x		x	x			x	x	
Companies					x	x	x		
Construction process									
Educational assessment				x	x	x			
Competence assessment	x								
Psychological tests									x
Test product and use									
Educational assessment		x	(x)*				x	x	x
Competence assessment							x		
Psychological tests			x					x	

*Although COTAN's focus lies on psychologcial tests, the system is also used to evaluate educational assessments.

Table 6.1 displays all of the systems for comparison and the three aspects that they are compared on. The object of evaluation is divided into two main objects: construction process and test product and use. Systems aimed at evaluating the

process tend to give guidelines for developing solid tests, whereas systems that focus on the test product and use are meant for auditing a fully developed test that is already in use. One element that stands out from this table is that the AEA-Europe system is multi-functional. That system aims to be a framework of standards that can be used in several different ways and for all sorts of educational assessments. In the remainder of this section, we will compare the systems in detail for each of the three aspects in Table 6.1.

Purpose

In our comparison, we distinguished four main purposes for the quality evaluation systems, guidelines, and standards. First, we looked at systems aiming to guide test development. These systems try to help the test developer in constructing a sound test. Both the ETS standards and the Cambridge approach are meant to guide test development. Another purpose is to help users apply tests properly and to make them aware of the risks when they do not follow protocol. One example of a system that has the purpose of helping users understand the interpretation of test scores is the ITC document that has guidelines for test use. Some systems are meant for self-evaluation by the test constructors, to help them identify the strong and weak points of their assessment; Baartman formulated criteria for this specific purpose. Finally, we included systems meant for audit purposes. In this case, an external expert audits the quality of the test by means of an evaluation system, such as the COTAN system or the EFPA system.

Intended audience

The intended audience of the evaluation systems can be test specialists, teachers, users, or companies. However, most of the systems that we compare are developed for multiple audiences. The systems that have only one intended audience are the ITC document (teachers), the Cambridge approach and ETS standards (companies), and EFPA (test specialists). The COTAN and AERA systems are meant for test specialists as well as teachers. The JCTP standards are intended for both teachers and test users.

The object of evaluation

The definition of quality also varies across the different systems. Some systems focus on the construction process, while others focus on the fully developed test and its use. For example, COTAN focuses on the fully developed test product and not on the development process. ITC, however, intends to evaluate the

development process. At the same time, the type of test differs: JCTP focuses on classroom assessment, while Baartman focuses on competence assessment programs. The AEA-Europe framework of standards focuses on educational assessment in general, where COTAN aims at both psychological and educational tests.

Issues with the currently available systems

One problem with all of these evaluation systems is that quality is defined as a non-flexible construct. These systems provide criteria that should be met, while it is actually more appropriate to choose criteria that fit the intended use of the test. Doing this would also provide the possibility of weighing the criteria according to the purpose of the test. This might solve the problem of having to create a new evaluation system for every type of test. Once the purpose of the test defines the selected criteria, we can also evaluate several types of tests with the system.

Another problem with these evaluation systems is the process of evaluating the tests. To evaluate a test as part of an external audit, one needs to look through all of the testing materials and supporting documents that include the results of the trial administrations of the test, validation studies, and other evidence that is considered relevant for the audit. However, it depends on the auditor whether all of the evidence is found. Moreover, going through all of these documents is not a very time efficient way to evaluate tests, and a lot of both content and methodology expertise is needed to evaluate a complete assessment (Wools, Sanders, Eggen, Baartman, & Roelofs, 2011). When a new evaluation system can make classifying evidence a task for test developers, the auditors only have to look through the relevant evidence. And when all of the evidence is structured in advance, it is also possible to give a part of the test to an auditor who knows the content and another part to an auditor who specializes in methodology.

These issues are addressed as principles in the outline of the proposed evaluation system. The design of the new system tries to gain from existing evaluation systems as well. In the remainder of this paper, the design of the new system is described.

Section 2 - Principles of the new evaluation system

In the new evaluation system, quality is defined as the degree to which something is useful for its intended purpose. In testing and assessment practice, the variety of intended purposes is very large and, furthermore, the solutions chosen to reach those purposes are endless. And, when quality is defined as being dependent on the purpose of a test, it seems hard, or even impossible, to develop an evaluation system with fixed criteria that are suitable for all possible tests and assessments. Therefore, we do not aim to develop the right set of criteria that can be used to evaluate all possible tests. The main idea behind this system is for it to be used to build an argument that helps test developers to show that a test or assessment is sufficiently useful for its intended purpose. To build this argument, evidence is needed to convince the public of the test's usability. This evidence is established, collected, and presented during the test's development process.
The argument-based approach to quality is derived from the argument-based approach to validation, as described by Kane (2004; 2006). The remainder of this section extracts the argument-based approach to quality into the underlying principles of the new evaluation system. As a starting point for the specification of the principles, the purpose of the system is addressed.

Purpose

The purpose of the system is to evaluate the quality of tests and assessments on several occasions during the construction of a test. It might be used during the development stage to indicate weak spots that need attention or adaptation, or utilized to point out aspects that are in need of evidence in order to enhance the plausibility of the argument that is being built. When the development stage is finished, the system also needs to facilitate an external evaluation of the test. The criteria used are derived from existing evaluation systems, and may be chosen or combined based on the purpose of the test or the purpose of the evaluation.

Content

As mentioned before, quality is defined as the degree to which something is useful for its purpose. By taking an argument-based approach to quality, it is possible to interpret quality as an integral entity instead of a combination of isolated elements. This entails the possibility of an assessment to compensate for weaker points with strong points. Furthermore, this view does justice to the

fact that all aspects of an assessment are linked and cannot be evaluated without considering the others.
This view also implies that the instrument that is used to assess students and to generate scores cannot be evaluated without considering the use of these scores. In an argument-based approach to quality, the use of the scores, or the decision that is made based upon the scores, guides the test developer in determining the appropriate quality standards. This means that, on one hand, the intended decision resulting from a test is the main determiner in choosing the criteria that are necessary to evaluate the appropriateness of the test. On the other hand, the degree to which the test must comply with the standards is also based upon the intended decision. For a high-stakes certification exam that consists of 40 multiple-choice items, reliability, IRT model-fit, and validation by means of an external criterion might be more appropriate than any coefficient of inter-rater reliability. Whereas, in a selection procedure where two assessors are interviewing their own groups of students, inter-rater reliability and comparability seem to be the most important aspects.

Process

According to the argument-based approach to quality, an argument is built and evidence is collected, selected, and presented according to the shape of the argument. By selecting and presenting the appropriate evidence, the evaluation is prepared during the test construction phase. Once the (external) audit starts, the auditor does not need to go through all of the available material, but only investigates the evidence that is presented according to the structure of the argument. This not only makes the evaluation process more manageable for the auditor, but also enhances the comparability of the ratings of different auditors, because they all took the same evidence into account. Another advantage of structuring the evidence before auditing is that different auditors with different competencies, for example, psychometricians and content experts, can evaluate the parts that they specialize in.

Relationship to other evaluation systems

One of the reasons to evaluate test quality is that it is necessary to decide whether the use of a certain test for an intended decision is justified. We would like to know whether a test is good enough for the stated purpose. An argument that is built and accompanied with evidence and that is evaluated as plausible is, unfortunately, not an answer to the question of whether a test is good enough. Therefore, the new evaluation system also includes other

evaluation systems' criteria that do lead to a result that states whether a test is good enough. These criteria are built into the system in such a way that, once the evidence is structured in the different elements of the argument, the criteria will appear in clusters that match the order of the argument. The order of the criteria is different from the order in the original evaluation systems, but once every criterion is answered, the results will be presented according to the elements of the original evaluation systems. For example, COTAN's criteria are clustered differently, but the evaluation results will be presented in the seven categories that are distinguished by COTAN.

Section 3 - Design of the new evaluation system

The new evaluation system will be a computer application that consists of several modules. These modules are: design, evidence, evaluation, and report. The application is designed for use during the test development process, but can also be used for the evaluation of existing tests. However, once the existing tests are evaluated, the test constructors have to prepare the evaluation by designing and structuring the argument.

Design module

This module delivers the outline of an argument. Therefore, several steps need to be taken. To make sure a user will complete all of the necessary fields, this module is wizard based. It starts by posing questions about the characteristics of the test. Once the general information about the test is collected, the assumptions and inferences that underlie the quality argument are specified. To build the argument, first the focus will be on the shape of the argument. The amount of inferences that need to be specified depends, for example, on the purpose of the test. When the shape of the argument and the characteristics of the test are known, the actual building of the argument starts. For every inference, the underlying assumptions are described. Furthermore, possible counter arguments are also made explicit.

Evidence module

The evidence module consists of two parts. First, it facilitates the storage and structuring of the sources of evidence. In this module, a user can upload documents, graphical representations, research reports, or test materials. For every document that is uploaded, it is possible to enter a short description and

to add tags. These tags can be used in the evaluation module to help auditors select the right sources of evidence.
Second, it focuses on structuring and classifying evidence. Evidence can be selected and added to the inferences that are specified in the design module. Graphics show which inferences are backed up with evidence so that the user can see which inferences need more evidence.

Evaluation module

This module is designed to facilitate the evaluation process by combining the information given in the design and evidence module. Therefore, there are two main parts within this module: prepare and evaluate.
Within the prepare section, the test developer can choose the evaluation system that will judge the test and the argument can be reviewed. The evaluate section shows the specified argument and the uploaded evidence. Furthermore, the criteria, questions, or aspects from the chosen evaluation system are shown with every inference. An auditor can go through the inferences and evaluate the quality of the evidence based upon the given criteria. Only the evidence that is a part of an inference is shown, therefore, the auditor does not need to look for the appropriate evidence.

Report module

The report module can be used to retrieve the results of the evaluation. It can also be used to print parts of the argument or the accompanying evidence, for example, to construct a test manual that incorporates all of the evidence and that is structured according to the specified argument.

Conclusion

This paper outlines a new evaluation system for the quality of tests, assessments, and exams. The new evaluation system will be developed as part of a study that will be shaped according to the principles of design research (Plomp, 2007) and will be finished in the summer of 2013. This system will incorporate an argument-based approach to quality, and we will suggest a computer application that can be used to gather, structure, and evaluate the evidence of quality. By explicitly using sources of evidence that are created during the different phases of the test development process, this new

evaluation system will bring new awareness of quality issues to everyone involved in test development.

The argument-based approach to quality is based upon a theory used in validation practice. This gives us the opportunity to look at quality in a more comprehensive way. From here, it is also possible to evaluate and weigh evidence in respect to the purpose of the test. Furthermore, where other evaluation systems focus on the end product of the test development phase (the test), this new evaluation system bridges the development efforts to the end product.

This new evaluation system will, however, also include existing quality criteria, which makes an evaluation according to the existing evaluation systems still possible. In conclusion, the proposed evaluation system will allow us to evaluate test quality in a flexible and comprehensive way, and gives us a conclusion about test quality from other evaluation systems at the same time. Could this be the system that combines the best of both worlds?

References

AEA-Europe. (2012). *European framework of standards for educational assessment.* Retrieved from http://www.aea-europe.net/index.php/professional-development/standards-for-educational-assessment

American Educational Research Association (AERA). American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999), *Standards for educational and psychological testing.* Washington: American Psychological Association.

Baartman, L., Bastiaens, T., Kirschner, P., & van der Vleuten, C. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*(2), 153–170.

Bartram, D. (2001). The development of international guidelines on test use: The international test commission project. *International Journal of Testing, 1*(1), 33–53.

Cambridge Assessment. (2009). *The Cambridge approach. Principles for designing, administering and evaluating assessment.* Cambridge: Cambridge Assessment.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education. Washington, DC: Joint Committee on Testing Practices.*

Educational Testing Service (ETS). (2002). *ETS standards for quality and fairness.* Princeton, NJ: ET.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*, 135–170.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Keuning, J. (2004). De ontwikkeling van een beoordelingssysteem voor het beoordelen van "Computer Based Tests." *POK Memorandum 2004-1.* Arnhem: Citogroep.

Lindley, P., Bartram, D., & Kennedy, N. (2004). *EFPA review model for the description and evaluation of psychological tests.* Retrieved from

Plomp, T. (2007). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 9–35). Enschede, Nederland: SLO.

Sanders, P. (2011). *Beoordelingsinstrument voor de kwaliteit van examens.* Enschede: RCEC.

Wools, S., Sanders, P., Eggen, T., Baartman, L., & Roelofs, E. (2011). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments. *Pedagogische Studiën, 88*, 23–40.

Wools, S., Sanders, P., & Roelofs, E. (2007). *Beoordelingsinstrument: Kwaliteit van competentie assessment.* Arnhem: Cito.

Chapter 7

An evaluation system with an argument-based approach to test quality

Abstract

To evaluate the quality of tests and assessments, we suggest to weigh criteria, depending on the intended interpretation and use of the assessment. This paper presents an online evaluation system for the quality of tests: the Quality Evaluation Application (QEA). The QEA was developed as part of a design-based research project. Its theoretical foundation originates within validation theories, and it incorporates an argument-based approach to quality. Design principles are formulated from these theoretical foundations. These principles are subsequently translated into a software prototype. This paper describes the development of the software as well as two evaluation studies on whether the initial design principles were met. The paper concludes with the assertion that the software is promising to foster discussion during test construction. Furthermore, the interface seems sufficiently useful to initiate a new phase in the research project whereby the software can be implemented for use by practitioners.

Keywords: quality, evaluation, software, validation

In all levels of education, students are required to take tests and assessments to demonstrate their ability. In the context of high-stakes exams and assessments, the importance of good quality decisions is clear. However, when assessment results are used for low-stakes decisions or, for example, for formative purposes, they need to be valid *and* reliable (AERA, APA, NCME, 1999). In other words, despite the stakes of an exam, the results need to be appropriate for their intended use. This can only be guaranteed when the assessment instruments used to assess students are sound. There are several evaluation systems and standards for the purpose of evaluating test quality (e.g., Bartram, 2001; Evers, Lucassen, Meijer, Sijtsma, 2010; AEA-E, 2012). These systems, however, are very specific: they tend to focus on one specific type of test or test use and aim to evaluate single assessments instead of assessment programs (Wools, 2012). Compared to auditing systems, standards are often more broadly defined, but they are aimed at guiding test developers during the development process and are not suited for external quality evaluations.
The purpose of this paper is to present a new online evaluation system for the quality of tests: the Quality Evaluation Application (QEA). The QEA was developed as part of a design-based research project (McKenny & Reeves, 2012). Its theoretical foundation originates within validation theories and will be described in the first part of this paper. Design principles are formulated from this theoretical framework, and a prototype of the QEA is built and evaluated. This iterative process is described chronologically and includes two studies evaluating the QEA.

Theoretical Framework

Since tests and assessments are used to make important decisions about students, the quality of these instruments needs to be evaluated. The evaluation of quality can be part of the construction process or can be performed after the assessment is administered, for example, as part of an external audit. The criteria used for these evaluations differ according to the moment of evaluation, the type of test, and the purpose of the test. Since these criteria differ, several evaluation systems are available. These systems tend to focus on one specific type of test or test use, for example, computer-based tests, competence-based assessments, or psychological tests. Standards are often more broadly defined but are aimed at guiding test developers during the development process and are not specifically constructed for external quality evaluations. In comparisons between existing evaluation systems, we can differentiate between guidelines,

standards, and auditing systems. Guidelines suggest quality aspects that you *can* comply with. Standards are aspects of quality that you *should* comply with in order to develop sound and reliable tests. Auditing systems focus on evaluating a test and prescribe what quality aspects *must* be met to ensure minimal quality. These systems present norms alongside the criteria that should be met to obtain a positive outcome with the use of the evaluation system. In a comparison between several evaluation systems (Wools, 2012), it became clear that all such systems define quality as a non-flexible construct.

However, we often want to be able to weigh criteria and use different norms according to the purpose of a test. Is, for example, the same reliability required for a national test used to admit students to higher education as for a test that is used by one teacher to evaluate whether students obtained the learning goals of a particular lesson? We probably want to be able to have different norms on quality for these various purposes. A new evaluation system should therefore facilitate different quality criteria and flexible norms for different test purposes.

In the new evaluation system presented here, quality is defined as the degree to which an assessment instrument is appropriate for its intended purpose. In testing and assessment practice, there is considerable variety in relation to intended purposes. The number of solutions chosen to reach those purposes is also endless. Moreover, when quality is defined as dependent on the purpose of a test, it is difficult, or even impossible, to develop an evaluation system with fixed criteria and norms that are suitable for all possible tests and assessments. Therefore, we do not aim to develop the right set of criteria for the evaluation of all possible tests. The central idea behind this system is for it to be used to build an argument that helps test developers demonstrate that a test or assessment is sufficiently useful for its intended purpose. To build this argument, evidence is needed to convince all stakeholders of a test's usability. This evidence is established, collected, and presented during the test's development process.

The argument-based approach to quality is derived from the argument-based approach to validation, as described by Kane (2006, 2013), and consists of three steps (Figure 7.1):

1. Building an interpretation and use argument (IUA)
2. Building a validity argument
3. Evaluation of the arguments and available evidence

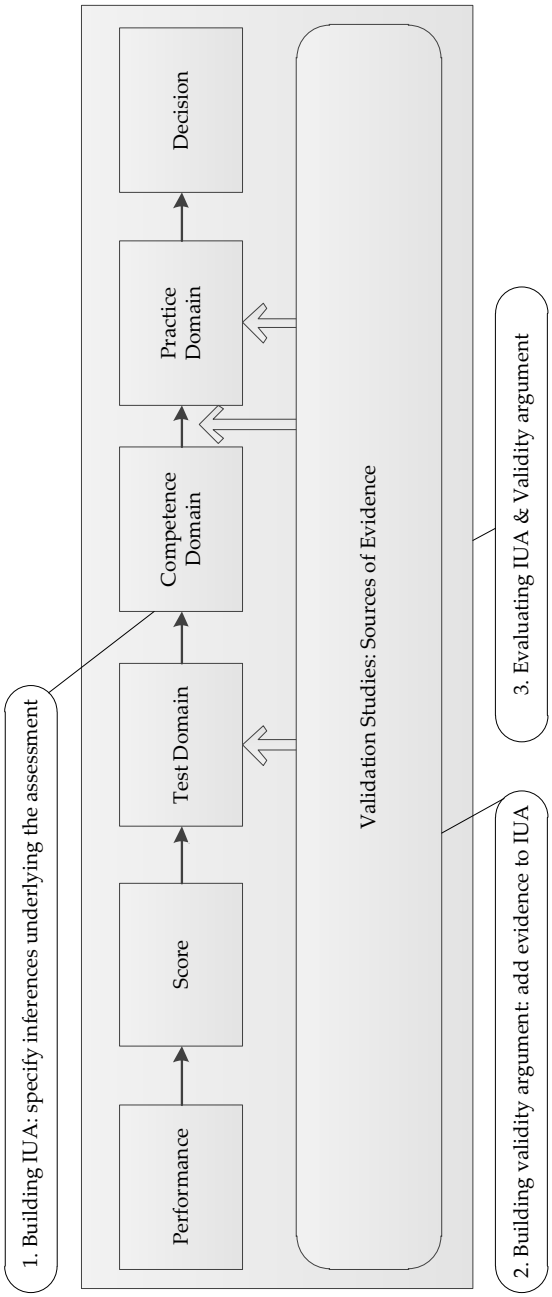


Figure 7.1: Schematic overview of three steps: building IUA, building validity argument, evaluation of both arguments.

The first two steps in Figure 7.1 correspond with the two stages of the original argument-based approach to validation (ABA): building an interpretation and use argument (IUA) and building a validity argument (Kane, 2013). Within the IUA, test developers specify the inferences that are drawn when interpretations of students' performances lead to decisions that match the intended purpose of the test. These inferences are represented by the arrows in Figure 7.1. Each arrow comprises an argument that is shaped according to the Toulmin model (Toulmin, 1958, 2003), which supports our reasoning from one domain to the next (squares in Figure 7.1). The Toulmin model prescribes the elements that can be used in this reasoning: warrants, backings, and rebuttals. When all inferences drawn within an assessment situation are made explicit according to the Toulmin model, an in-depth evaluation of the plausibility of all parts of the assessment is necessary. This in-depth evaluation takes place in the second stage of the argument-based approach. Evidence is then collected and combined into a validity argument to support or reject the inferences that are specified within the interpretive argument. When the argument-based approach is extended from validation to quality, an additional stage is added (Wools, Eggen, & Sanders, 2010). This third step in Figure 7.1 yields a final evaluation of quality. This is performed by an auditor who reviews the constructed arguments and the collected sources of evidence and decides whether it is plausible that the assessment is sufficiently useful for its intended purpose.

It is time-consuming to build these arguments, specify the inferences, and classify evidence. Moreover, a significant amount of knowledge of the specific model is necessary to be able to use this particular approach. In validity research, several authors stress the need to support practitioners in using the ABA. Some (Sireci, 2009, 2013) choose an approach whereby the ABA is adapted to a simpler form. Others (Llosa, 2008; Wools, Eggen, & Sanders, 2010) exemplify the ABA by validating actual assessments as a way of developing examples that practitioners can follow. Both approaches, which aim to support practitioners in using the ABA, have disadvantages. Simplifying the ABA causes the approach to lose some of its depth, usability, and nuance. These simplified approaches stress the need to gather multiple sources of validity evidence but do not offer guidance on choosing the evidence that is most needed, given the intended interpretation and use of the test. This means that these approaches simply encourage users to gather all sources of evidence or at least evidence in each predefined category. In the end, this does not lessen the burden for practitioners or help them understand the need to prioritize the sources of evidence that support the weakest claims. The other approach,

exemplifying the ABA, can clarify some issues in relation to the ABA, but it does not support practitioners in building their own arguments. Thus, following this approach, practitioners are still not supported in their actual efforts for validation.

This paper presents a software that aims to make the argument-based approach more accessible and useful for practitioners without altering the depth and complexity of the theory and by doing more than simply exemplifying it. The software guides users in the process of building quality arguments and evaluating these arguments. During this process, users are visually supported in their efforts.

The developed software – the Quality Evaluation Application (QEA) – was developed within a design-based research project (McKenny & Reeves, 2012). This iterative method includes building a theoretical framework into design specifications, extending these specifications into a prototype, and evaluating whether the specifications are met within the prototype. In the project described in this paper, we performed two cycles of this process. A prototype (QEA 1.0) was built according to the design specifications derived from the theoretical framework described by Wools (2012). This prototype was evaluated during Study 1. The results of Study 1 were then interpreted against the design principles, and a new design was proposed. This new design was implemented in a new version of the software: QEA 2.0. This new prototype was subsequently evaluated in Study 2. The results of these studies were compared to evaluate whether the adjustments improved the software.

These cycles are described chronologically. First, the design principles are introduced, and the first prototype of the software is subsequently presented. Following this, the methods and results of the first evaluation study are described. The subsequent section presents a new version of the developed software based on the results of Study 1. This new version was evaluated in Study 2, and the methods and results of this study are presented after the description of the adjusted software. In the final section, conclusions are drawn on the extent to which the software fits the initial design principles.

Development of QEA 1.0

Design principles

As a basis for the development of the prototype, four design principles were formulated. These principles were translated into a software prototype that

aims to help users in applying the argument-based approach to quality. In this section, the general functionality of the software is described in relation to the four design principles.

1. The system can be used during several stages of the test development process: it is not only suitable to guide test development, but can also be used as an instrument for internal or external audits.

Users are able to use the software during test construction to help them structure design choices and elucidate underlying claims and inferences that follow from these choices. Furthermore, the software can be used to store and classify evidence generated during the test construction process. Since these sources of evidence are connected to the inferences that they support or reject, the classification of evidence also prepares the evaluation of the assessment for auditing.

2. The system defines quality as the extent to which something is useful for its purpose and, therefore, incorporates an argument-based approach to quality.

The QEA supports users in building arguments according to the rationale of the argument-based approach. Furthermore, an evaluation system that matches this approach is incorporated to evaluate the quality of assessments by using flexible quality criteria that fit the intended use of the assessments.

3. The auditing process is simplified, and auditors' judgments should become more comparable.

Due to the preparation of the auditing process during the construction phase, auditors do not need to look for evidence through an unstructured collection of sources of evidence, but only consider the sources of evidence connected to the inference they need to evaluate. This way, differences in judgment on quality do not depend on the sources of evidence found by the auditor but on the weight that auditors put on the evidence. Furthermore, it is possible to assign specific inferences to specific auditors to make sure the expertise of auditors matches the sources of evidence. An advantage of this approach is that those responsible for combining the judgments of different auditors can gain greater insight into the elements considered by these auditors.

4. The system should include other evaluation systems to prevent it from being just another evaluation system.

Several evaluation systems are included to facilitate the use of different norms or standards in the auditing process. In the current prototype, two evaluation systems were incorporated: a system that fits the argument-based approach (Wools, Eggen, & Sanders, 2010) and a Dutch evaluation system (Evers et al., 2010).

QEA 1.0

This section describes the prototype in greater detail, which is build according to the design principles. The first version of the Quality Evaluation Application (QEA 1.0) consists of three modules. The first is a design module, the second, an evidence module, and the third, an evaluation module. The design and evidence modules are targeted at test developers while the evaluation module is aimed at auditors. Furthermore, all modules are usually used in chronological order: design, evidence, evaluation. However, this order is not enforced by the software. Therefore, it is possible to switch between the design and evidence modules during a test development process. When the application is used for internal, formative audits, it is also possible to keep track of the quality during the construction process to ensure that changes in the design can still be made.
Next to the three core modules in the software, a user and project management module is incorporated. In this module, projects and users can be created, edited, and deleted. In this part of the software, different roles are identified. Test developers can create projects and use the design and evidence module. Auditors can only evaluate assigned projects in the evaluation module but cannot enter the other modules. Finally, "Admins" can use all modules and can create users and assign them to roles and projects. The remainder of this section describes the three core modules.

Design module

Figure 7.2 displays the design module used to enter the characteristics of a test and to build the outline of the quality argument. At the top of this figure, an argument that includes domains and inferences is displayed. Domains are represented as squares, and inferences are the connecting lines between the domains. For every inference in the design module, an argument shaped according to Toulmin's model can be created. An example of such an argument is shown at the bottom of Figure 7.2. It is possible to add or remove elements of the Toulmin model to build it according to specification.

When a user enters the design module, a basic form of an IUA, as specified by the argument-based approach, is displayed. For every domain, a user can enter a title and description. QEA 1.0 facilitates quality arguments aimed at one test, but it can also be used to build arguments that fit assessment programs (Wools, Eggen, & Béguin, submitted). For this latter use, multiple domains under performance, score, test domain, and competence domain can be added. The theory relating to the validation of assessment programs restricts adding multiple practice domains and decisions; therefore, the software does not support this. When users want to include multiple decisions, they need to build multiple arguments and create multiple projects.

Evidence module

The evidence module is used to upload and classify different sources of evidence. Once the argument is shaped and filled within the design module, users can upload different sources of evidence. There are no restrictions on the file types that users can upload. For every file that is uploaded, users can enter a short description in relation to the content or how this file might be used as evidence for a particular claim. When files are uploaded, classification is done by connecting the uploaded file to a domain or to a single element of a Toulmin argument. This way, evidence relating to a specific inference is connected in the software to that particular inference. Figure 7.3 shows a screenshot of the evidence module where one source of evidence is connected to a Test domain. More specifically, a test matrix is uploaded (Test matrix.xlsx) and connected to the Test domain, Math.

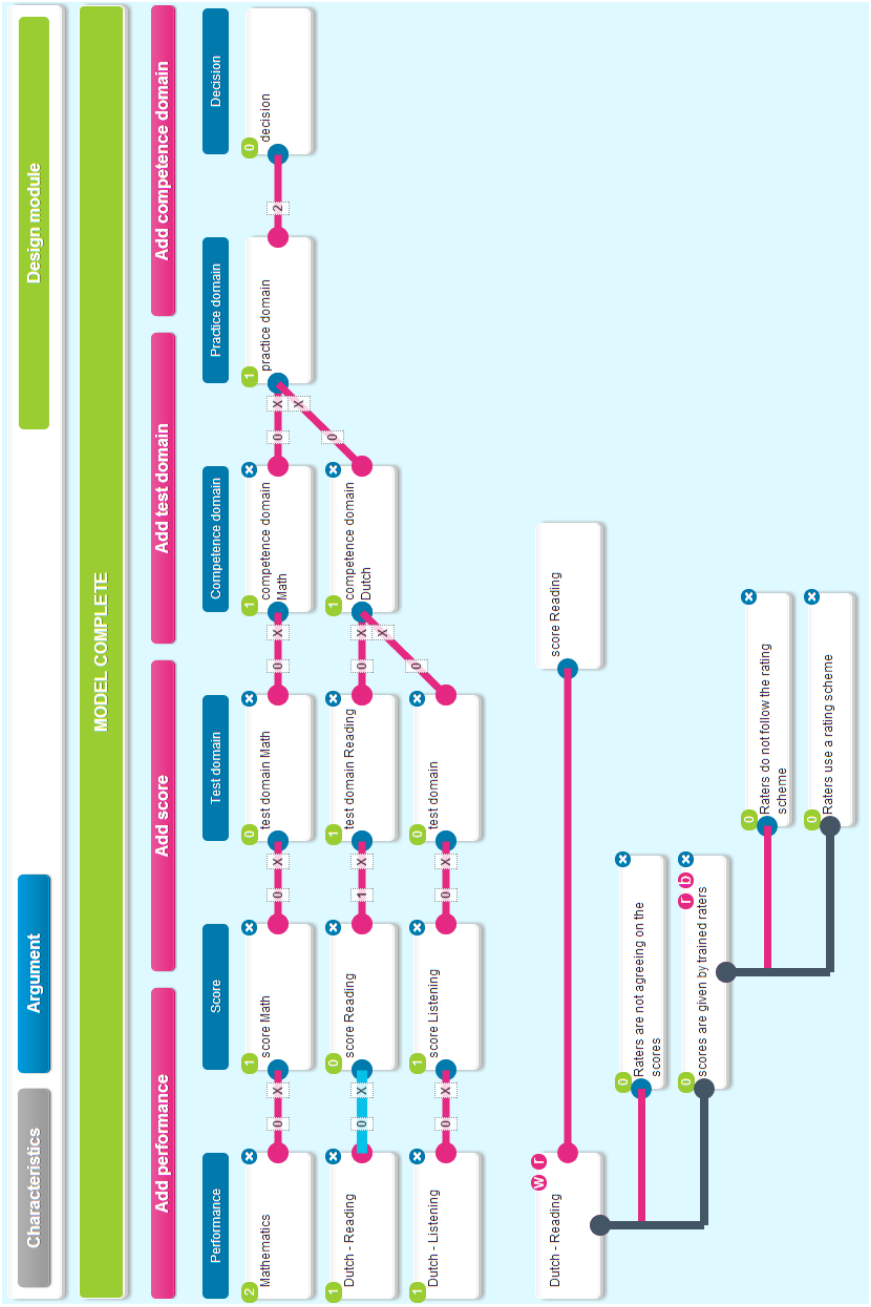


Figure 7.2: Design module in QEA 1.0 – used to build arguments

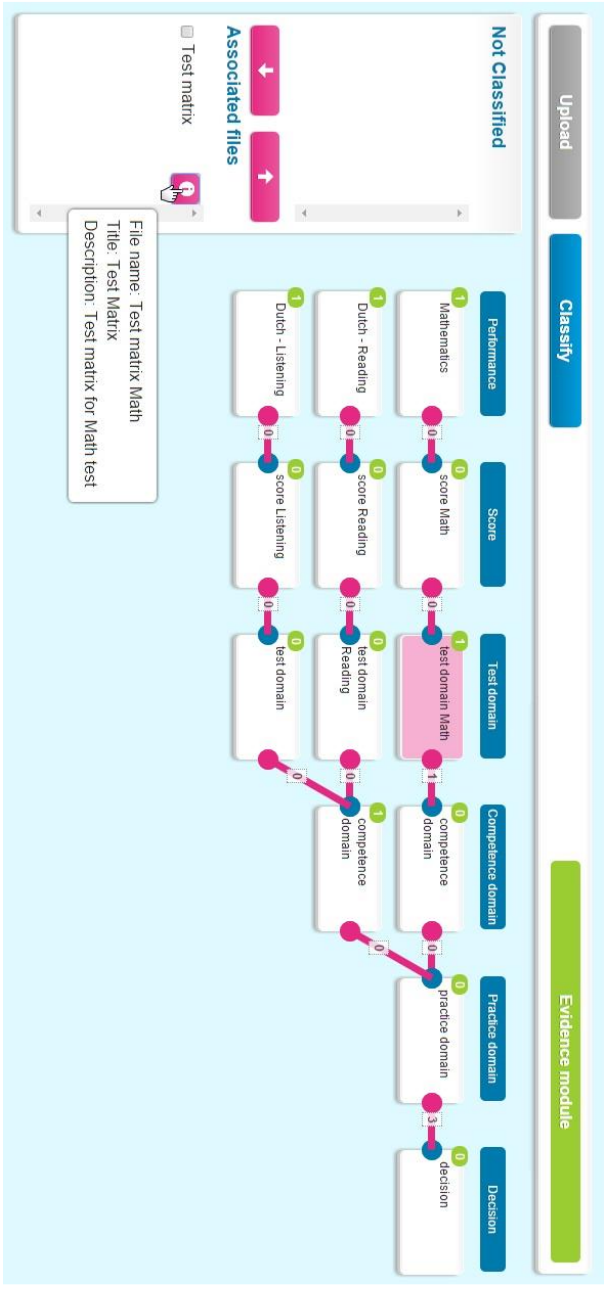


Figure 7.3: Evidence module QEA 1.0 – used to upload and classify evidence

Evaluation module

Within the evaluation module, the sources of evidence are presented (as they were classified in the evidence module) and reviewed according to criteria provided in the software. Auditors are able to open the sources of evidence related to an inference. Once they have evaluated the evidence, they can enter their judgment. The specific criteria or questions posed to auditors depends on the evaluation system used. The current version of the QEA incorporates two evaluation systems, the first of which matches the argument-based approach to quality (Wools, Eggen, & Sanders, 2010). Figure 7.4 displays a screenshot of this particular evaluation system. The major part of the screen displays the argument constructed by the test developer and the underlying Toulmin arguments. In the upper left-hand corner, the sources of evidence connected to an inference are displayed. Based on these sources of evidence, an auditor can decide on every element of a Toulmin argument as to whether it is "accepted," "rejected," or "unclear." To indicate their judgement, auditors need to press the button reflecting their choice for every element, as illustrated in Figure 7.4. Once all the domains and inferences are evaluated, a result is calculated based on scoring rules incorporated in the evaluation system.

The second evaluation system in QEA 1.0 is the COTAN system (Evers et al., 2010). This Dutch evaluation system for the quality of educational and psychological tests includes 83 questions to be answered and leads to a decision on quality in terms of "Insufficient," "Sufficient," and "Good" for seven categories. These categories are: Theoretical Basis of the Test, Quality of Test Materials (Paper-based), Quality of Test Materials (Computer-based), Norms, Reliability, Construct validity, and Criterion validity. The questions within the system were restructured according to the domains and inferences of the basic form of the argument-based approach. All questions were classified under an inference or domain. This classification was done in a similar way as in a previous study in which sources of evidence were classified into the domains and inferences of the argument-based approach (Wools & Eggen, submitted). To calculate the results of an evaluation, the scoring rule of COTAN is translated and built into a software. The results of this COTAN evaluation system are reported according to the original categories. In contrast, the results of the argument-based approach to quality are presented on the basis of the argument that a test developer has specified in the other modules.

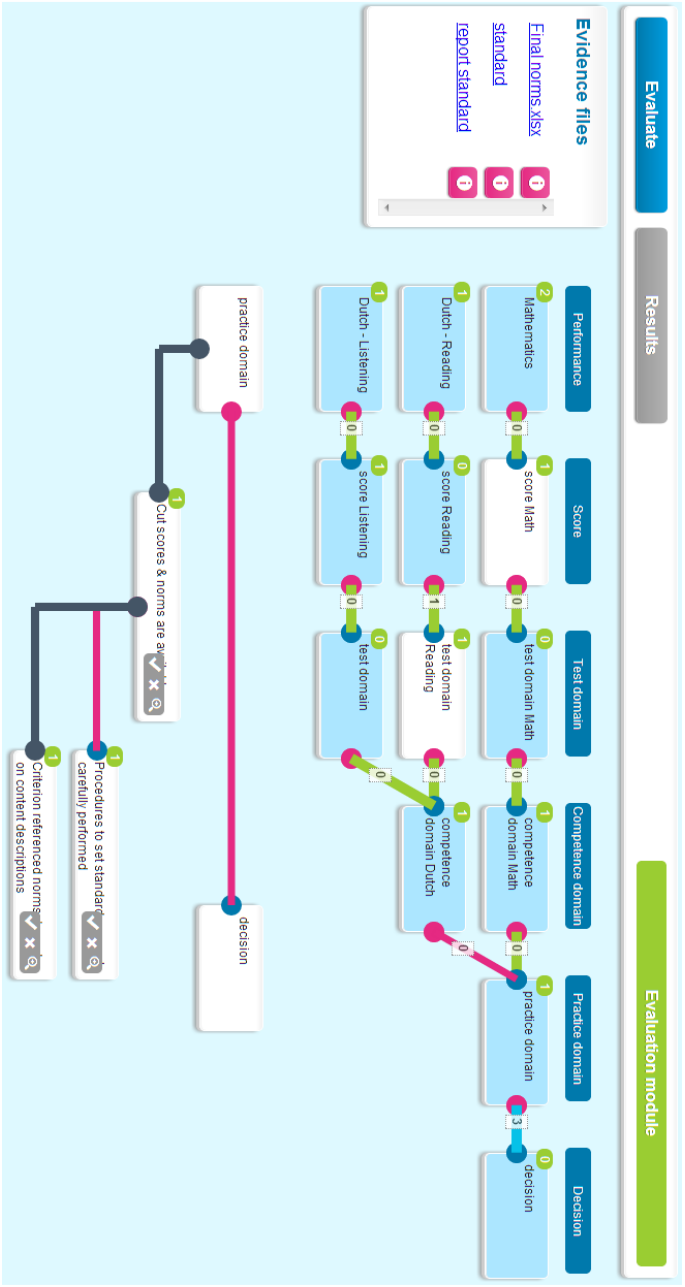


Figure 7.4: Evaluation module of the QEA 1.0 system related to the argument-based approach to quality – used to evaluate the quality of tests

Study 1

Once QEA 1.0 was finished, a study was conducted to evaluate whether the initial design principles were met. However, the fourth design principle (the software includes multiple evaluation systems) was already accounted for during the development the software. Therefore, this study focuses on evaluating the first three design principles. To evaluate whether these three design principles were met, a research question was generated for every principle:

1. Does the software support users in building arguments, selecting and classifying sources of evidence, and evaluating their assessments?
2. Are stakeholders agreeing with the theoretical framework and underlying design choices made while implementing the argument-based approach to quality into the QEA?
3. Is the interface of the QEA intuitive and easy to use so that the software itself does not complicate the auditing process?

To answer these research questions, a focus group was asked to respond to the prototype of QEA 1.0. The procedure, participants, and materials are described in the next section.

Methods Study 1

Procedure

A focus group was organized during the bi-annual conference of the European Association of Research in Learning and Instruction (EARLI) in Munich (Wools, 2013). Conference participants voluntarily attended a 90-minute workshop, which was part of the conference and open to all attendees – mainly educational researchers. At the workshop, a demonstration of the software was alternated with filling out questionnaires about QEA 1.0. During the session, information was given on specific parts of QEA 1.0 in three rounds, and a questionnaire was completed in every round. The aim during the rounds was to give just enough information for participants to be able to fill the questionnaires, but not too much so that they could respond spontaneously.

Materials

To accommodate the different backgrounds of the participants of the workshop, two different types of questionnaires were used (Appendix I). They addressed

either the theoretical model underlying QEA 1.0 (T-Questionnaires) or its interface and usability (I-Questionnaires). Every questionnaire consisted of three parts that fit the three modules in the software.
The T-Questionnaires related to the first and second research questions and included items about the theoretical model incorporated in the software. It also addressed specific choices that were made in implementing the original theoretical framework into the adapted version of the software. A question addressed, for example, the limitation of one practice domain and decision. Another question was related to the amount of evidence needed to evaluate an interpretive argument. All questions were coded as correct or incorrect when applicable, or more extended answers were recoded into more comprehensive categories.
The I-Questionnaires focused on the interface and usability of QEA 1.0 and aimed to collect data to answer the third research question. The questions in the questionnaire addressed certain elements that could be confusing or unclear in the interface design. Therefore, respondents were asked, for example, to explain what they thought certain actions in the software would result in; or they were asked to write down the actions they thought were necessary to reach a certain goal in the software. All the questions were coded as correct or incorrect.

Participants

The number of respondents differed for the three rounds since the questionnaires were distributed upon request (Theoretical, Interface, or both), and some respondents did not have enough time to complete every questionnaire. Table 7.1 displays the number of respondents for each round and questionnaire. In total, 26 participants handed in one or more questionnaires.

Table 7.1: Number of respondents per round and questionnaire

Round	Questionnaire T	Questionnaire I
1	18	16
2	8	13
3	5	9

Results Study 1

Theoretical Framework Questionnaires

The results of the theoretical framework questionnaire are grouped according to four themes that all answer the research question regarding the theoretical framework of QEA 1.0. The first two themes, "examples of sources of evidence"

and "evaluation of evidence," focus on the ability of participants to use the theoretical framework in a way that would be expected of test developers or auditors. The first research question is answered through these themes. The other two themes, "design principles" and "model restrictions," answer the second research question relating to the translation of the theoretical framework into the software.

1. Examples of sources of evidence

In the questionnaire, participants were asked to think of evidence that could be presented to support claims made within a presented example. Not all participants who completed this part of the questionnaire were able to answer this question. The examples of the sources of evidence that the participants provided are listed in Table 7.2. This table shows that for some claims, participants agreed on the evidence necessary to support them.

Table 7.2: Examples of evidence provided by participants in Study 1

Claim within scoring inference	**N**	**Examples of evidence**
Rebuttal 1: Raters are not agreeing on the scores	5	• rating form from two raters (3x) • inter-rater reliability study and results (2x)
Warrant: Scores are given by trained raters	5	• notes of discussion session among raters • description of training program (3x) • report on their working procedure • in-depth qualitative research on how raters use the scale
Rebuttal 2: Raters do not follow the rating scheme	3	• examples and number of occurrences of deviations from rating scheme • interview raters on their use of the rating scheme
Backing: Raters use a rating scheme	5	• the rating scheme that is used (3x) • an archive with all rating schemes (2x)

2. Evaluation of evidence

During the focus group discussion, participants were asked to evaluate the sources of evidence provided in the questionnaire. Three claims were presented with accompanying evidence. Participants were asked to choose, based on the provided evidence, whether they thought the claim should be accepted or rejected or whether further evidence was needed (unclear). Table 7.3 presents the choices of the five respondents who answered these questions. Remarkably, for none of the sources of evidence provided did raters agree on their judgment.

Table 7.3: Judgments of participants on presented evidence in Study 1

Participant ID	Warrant: Cut scores & norms are available	Rebuttal: Procedures to set standards are not carefully performed	Backing: Criterion-referenced norms based on content descriptions
20	Accepted	Rejected	Accepted
17	Accepted	Rejected	Accepted
10	Accepted	Rejected	Rejected
12	Unclear	Accepted	Accepted
11	Unclear	Rejected	Rejected

3. Design principles

Participants were asked to prioritize the four design principles mentioned earlier in this paper. A total of 17 participants answered the question. Table 7.4 displays the number of participants who prioritized a principle as the most important (1) to the least important (4). Most participants thought that Principle 2 (incorporate argument-based approach to quality) was most important and that Principle 4 (contains multiple evaluation systems) was the least important.

Table 7.4: Priority of design principles proposed in Study 1

Priority	Principle 1	Principle 2	Principle 3	Principle 4
# 1	3	9	3	2
# 2	6	2	7	2
# 3	5	5	5	2
# 4	3	1	2	11

Next to prioritizing the existing principles, participants were asked whether they felt a principle should be added. Five respondents gave an answer to this question. Two stressed the need for the system to be suitable for several types of education and its broad applicability. One indicated the importance of several users with different expertise to be supported by the QEA. Another participant indicated that it would be useful when the system was able to store multiple versions of an argument or test being evaluated. The final comment related to choosing between the simplicity of the model and guarding its comprehensiveness, the latter being the most important. Although participants added these principles it was decided that they all can be perceived as being included within the current principles.

4. Model restrictions

Respondents were asked to indicate whether they agreed with the design choice of only allowing for one practice domain and one decision per argument. Three out of eight participants who answered this question agreed with this choice. Five participants did however provide examples of where they thought multiple practice domains could be necessary. However, these examples could also be interpreted within an argument as different competence domains according to the definition used in the theoretical framework of the QEA. The following two quotes exemplify the responses of these participants:

> "Depends on the decision and on what 'level' the decision takes place. One decision, for e.g., a diploma (yes/no) at the end of the educ.program – but couldn't that include more practice domains, e.g., workplaces?" (Respondent 6)

> "It is not totally clear why you do not distinguish between different contexts. For example, differences in location, time, space (for example, a competent driver means something different in the US or in the Netherlands or on the moon or 100 years ago)" (Respondent 10; translated from Dutch to English)

Interface and Usability Questionnaires

The results of the I-Questionnaires were also grouped into four themes: (1) clarity of the modules, (2) building arguments, (3) indication of sources of evidence, and (4) evaluation module. All themes aimed to answer the third research question posed earlier in this paper.

5. Clarity of the modules

Participants were asked to describe the type of activity that could be performed within a module. Before answering this question, they were only briefly instructed about the general purpose of the QEA: a system to help users in building quality arguments to facilitate quality evaluations of assessments. The questions regarding the function of the modules were first open-ended and were then posed again in multiple choice format. The responses of 16 participants were scored as either correct or incorrect. Tables 7.5 and 7.6 display the number of participants who correctly indicated the function of the modules. From these frequencies it seems that it is rather hard for participants to recognize the function of modules based on their titles.

Table 7.5: Open-ended questions regarding the function of modules

Module	Correct	Incorrect	Missing
Design module	2	9	5
Evidence module	5	5	6
Evaluation module	3	8	5

Table 7.6: Multiple choice questions regarding the function of modules

Module	Correct	Incorrect	Missing
Design module	2	13	1
Evidence module	9	6	1
Evaluation module	12	3	1

6. Building arguments

Three questions were posed regarding the usability of the design module. The QEA aims to facilitate users in building quality arguments. Therefore, argument building must be quite intuitive in the software. In the questionnaire, participants were asked to answer questions that could demonstrate whether the software worked in the way they had anticipated. All answers were coded into correct or incorrect categories whereby "correct" meant that the indicated behavior in the software would lead to the desired result. "Incorrect" meant that users expected something different from the software or that their indicated behavior would not have the desired effect. Table 7.7 presents the questions posed and the number of participants who answered the questions correctly or incorrectly. These results indicate that the majority of participants would be able to use the software to build an argument.

Table 7.7: open-ended questions regarding building arguments

Question:	N	Correct	Incorrect
Q1: What do you think will happen if you press the button "Add test domain"?	13	12	1
Q2. You would like to add an additional rebuttal … What do you need to do to achieve this?	10	6	4
Q3: What do you think is necessary to build the complete argument?	11	9	2

7. Indication of sources of evidence

When sources of evidence are uploaded and classified within QEA 1.0, a number is shown to display the number of sources connected to an element of

the argument. However, the displayed numbers are not uniformly designed, as shown in Figure 7.5.

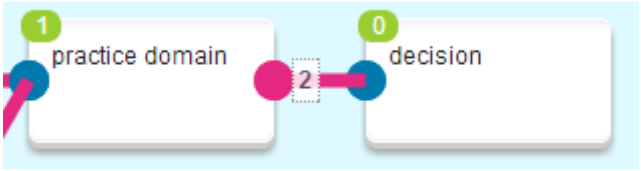


Figure 7.5: Close-up of indication of number of sources of evidence

For both the number in green at the corner of a domain and the number in the rectangle placed on an inference, participants were asked to indicate what it implied. The answers of 11 respondents were coded as correct or incorrect. For the numbers in green, six participants indicated the correct meaning while five were incorrect. However, for the numbers on the pink line, only two participants indicated the correct meaning, and five were incorrect. For these latter questions, three participants made no indications. Codes were also used to indicate whether participants had responded that both numbers had exactly the same meaning or that they had different meanings. All except one participant thought that the numbers had different meanings.

8. Evaluation module

The final part of the questionnaire concerned the evaluation module. The I-Questionnaire contained questions regarding the color scheme of the module. Participants were presented with the legend of the color scheme and were shown close-ups of the software. They were asked to indicate what a certain color meant in the software to ascertain whether they could understand the legend provided. This legend is also displayed in Figure 7.6. Five questions were posed, each of which was answered by nine respondents. However, of all 45 answers given, only six were correct.

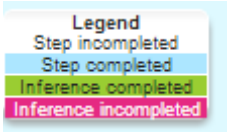


Figure 7.6: Legend of evaluation module

In this last part of the I-Questionnaire, a question was posed about the need to evaluate domains and inferences. It asked to indicate within an argument what element needed to be evaluated. This question was not only related to the

interface and usability of QEA 1.0 but was also very much related to the theoretical framework. From a theoretical perspective, QEA 1.0 is designed according to the principle that all elements need to be evaluated. However, a large majority of the participants who answered this question indicated that they felt that only elements with evidence need to be evaluated. This result should be evaluated from an interface perspective: is this result prompted by the design choices made in the interface? From a theoretical perspective, the question would be: is there a theoretical problem when only the elements supported with evidence are evaluated?

Redesigning the QEA

Conclusions were drawn from the results of Study 1 and resulted in suggestions for improving QEA 1.0. Table 7.8 displays the category of results, the conclusions drawn from these results, and the suggestion for improving the QEA. The results are presented in the same order described earlier in this paper.

Table 7.8: Suggestions for improving QEA 1.0

	Result	Conclusion	Possible solution
1	Example of sources of evidence	Users were able to identify sources of evidence necessary to support the claims within the interpretive argument. However, they indicated that they needed more support to do so.	Focus on simplifying the usability of the software. Once the interface becomes more straightforward, users might be supported to name the sources of evidence. This hypothesis has to be evaluated with a new interface.
2	Evaluation of evidence	Raters differed in their judgment of the evidence provided.	Once the QEA is developed, a study must be performed with real-life assessments that can be evaluated in relation to actual context and consequences for students.
3	Design principles	Respondents indicated that adding multiple evaluation systems is the least important aspect.	In altering the software, prioritize functionality and interface over adding more evaluation systems.
4	Model restrictions	Respondents agreed with the design choices made	Do not change the core functionality of the software.

	Result	Conclusion	Possible solution
		during the implementation of the argument-based approach in the QEA.	
5	Clarity of the modules	The titles of modules are still unclear; furthermore, users are not certain of the functionality within the modules.	Change the structure of the software to enhance the understandability of the modules.
6	Building arguments	Respondents were able to answer the questions relating to the correct building of arguments. However, the interface still has some inconsistencies that need to be solved.	Solve inconsistencies in the interface design.
7	Indication of sources of evidence	The current indications (in green and pink) are unclear. Users are not able to understand their meaning.	Change the indication of sources of evidence, and make it more consistent.
8	Evaluation module	The color scheme is unclear. Furthermore, users do not understand what needs to be evaluated.	Change the color scheme, and investigate whether a new interface solves the lack of clarity as to what needs to be evaluated.

To decide which suggestions should be incorporated in a new version of the QEA, the changes were weighed against the initial design principles. Respondents indicated that the fact that the system incorporated multiple evaluation systems was least important; therefore, it was decided that changes should focus on enhancing the usability of the software, as indicated in Suggestion 3. Furthermore, in Suggestion 4, it was concluded that respondents agreed with the design choices relating to the implementation of the argument-based approach to validation. Therefore, the new version of the QEA would differ in usability and interface design, but the core functionality of the software would not be adjusted.

Suggestions 1, 5, 6, 7, and 8 in Table 7.8 relate to interface aspects of QEA 1.0. These issues might be solved in a new design. Furthermore, it might be the case that when the suggestions relating to the interface are followed, the issues with the theoretical framework become less prominent. Therefore, the adaptations to

QEA 1.0 all relate to the interface of the software, and this version of the QEA was used to evaluate whether the remaining issues (Suggestion 1, 2, and 8) were resolved.

QEA 2.0

To improve the usability and clarity of the QEA, QEA 1.0 was evaluated against the First Principles of Interaction Design (Revised & Expanded) by Tognazzini (2014). These principles are of use for the design and implementation of effective interfaces. We classified the 22 principles into three groups: functionality, structure of software, and interface. The functionality-related principles are not within the scope of this improvement. When the principles within the structure of the software and interface were not met in the QEA 1.0 version, they were accounted for in QEA 2.0.
This resulted in two major thematic changes to QEA 1.0, each holding several small changes, though they are best explained in a broader sense.

Structure of QEA

The principles of interaction design relating to the structure of the software, which were addressed in QEA 2.0, are listed in Table 7.9. The descriptions added in Table 7.9 were used by Tognazzini (2014) to exemplify the principles and are meant to give a better understanding of them.

Table 7.9: Principles of interaction design related to the structure of the software (Tognazzini, 2014).

Principle	Description
Autonomy	Enable users to make their own decisions, even ones that are aesthetically poor or behaviorally less efficient.
Discoverability	Controls and other objects necessary for the successful use of the software should be visibly accessible to all.
Explorable interfaces	Offer users stable perceptual cues for a sense of "home."
Visible interfaces	Limit screen counts by using overlays.

The structure of the QEA was adjusted in such a way that users are able to navigate more freely through the application. The decision was made to combine the design and evidence modules into one test developer environment. Through the use of "tabs," navigation between the two initial modules was simplified. This change aims to make the building of an argument and the collecting of related evidence more intuitive and intends to approach this

process more iteratively. Furthermore, it was deemed less restrictive to users' autonomy when they want to build the interpretive and validity arguments at the same time.

The tabs are easily accessed by selecting the tab in the upper left-hand corner. In contrast, in QEA 1.0, a user was required to close the design module and open the evidence module through three different screens. This change in navigation is more direct and shows the test developer all possible modules at all times (principle of discoverability). Further, now that the module-selection screen has been deleted, the screen count (the number of different screens) is limited, as prescribed by the visible interfaces principle. In terms of the structure of the software, a new implementation means that users can return to the project-selection screen at all times by choosing the "home" button. This is in line with the principle of explorable interfaces.

Figure 7.7 shows a screenshot of the design tab in the test developer environment of QEA 2.0. This design tab replaces the design module from QEA 1.0. In this particular screenshot, the same example is shown as in Figure 7.2. Figure 7.8 shows an example of the evidence tab in the test developer environment.

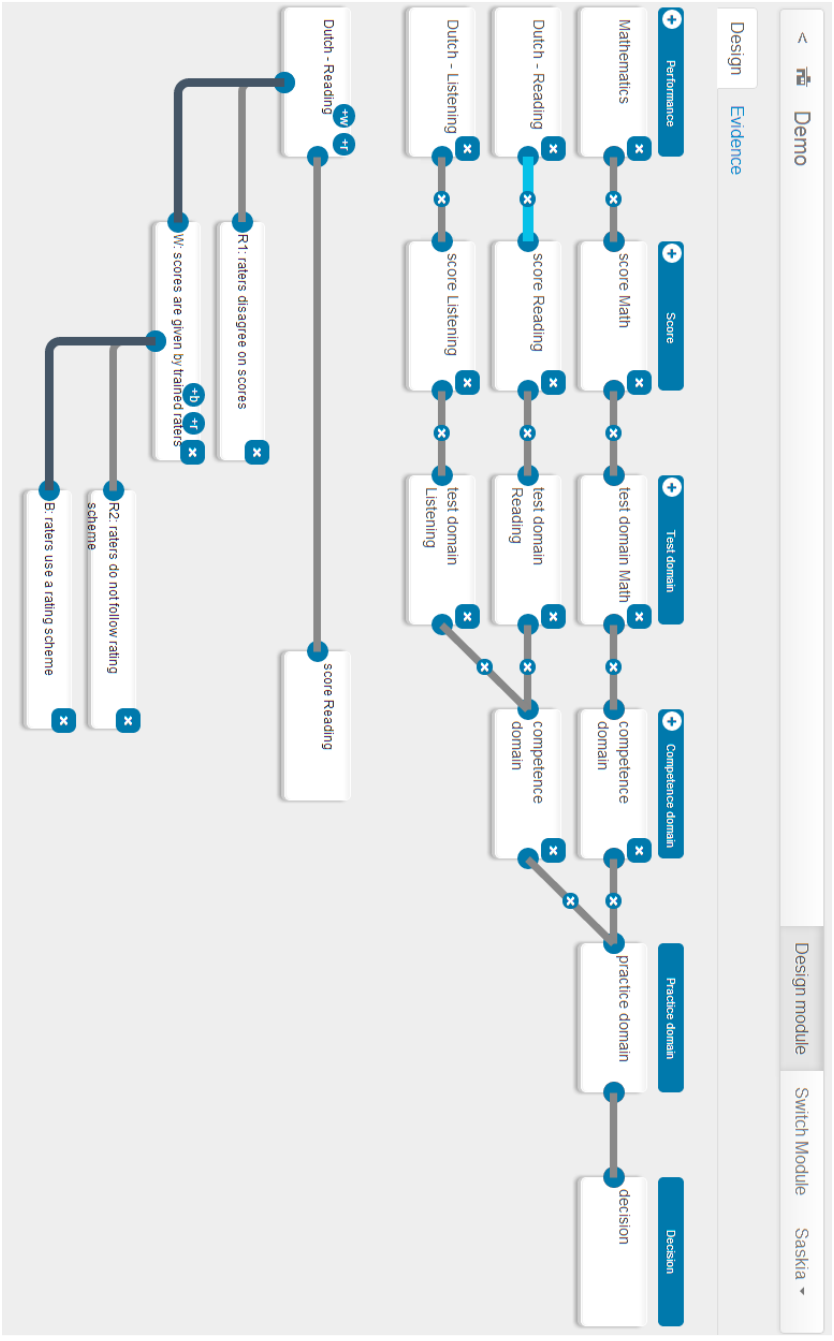


Figure 7.7: Design tab in QEA 2.0 – used to build arguments

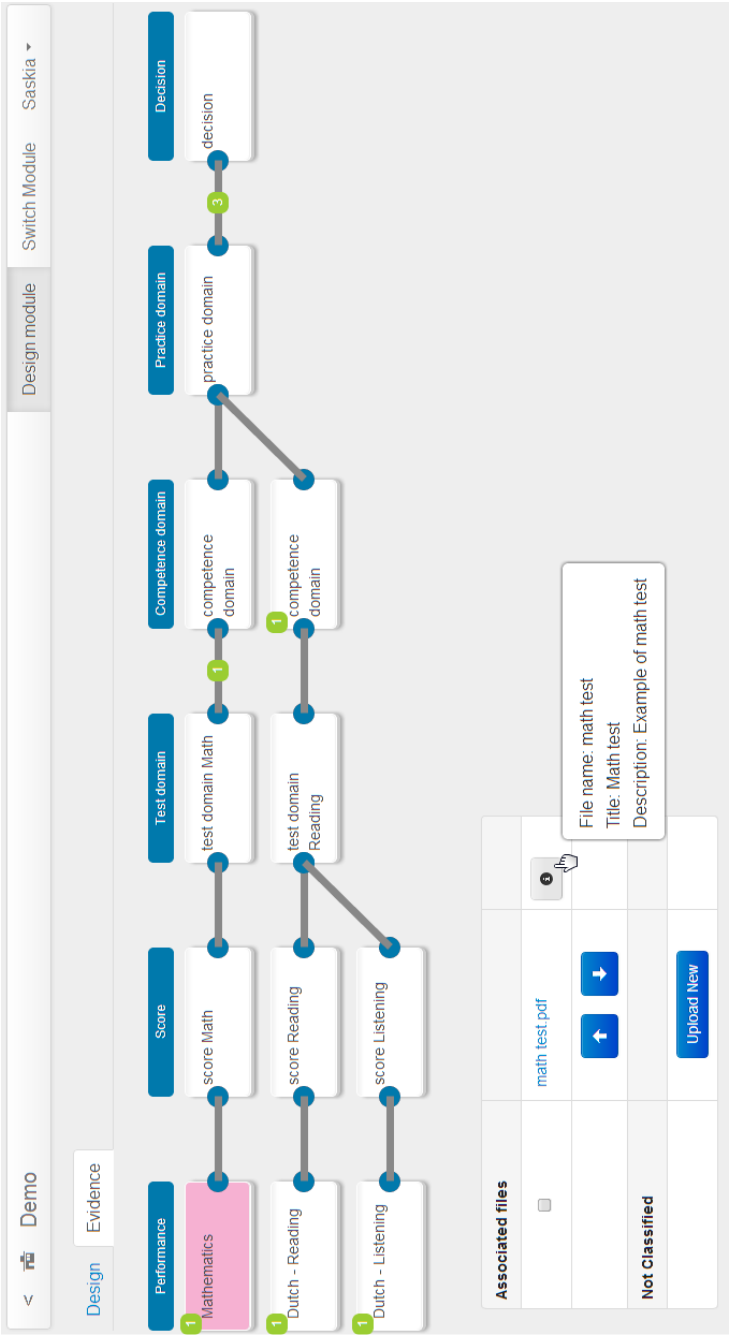


Figure 7.8: Evidence tab in QEA 2.0 – used to upload and classify sources of evidence

Interface design of QEA

A new interface design was developed for QEA 2.0. The principles of interaction design relating to the interface design were grouped together and listed in Table 7.10.

Table 7.10: Principles of interaction design related to the interface design (Tognazzini, 2014).

Principle	Description
Anticipation	Bring to the user all the information and tools needed for each step of the process.
Color	Any time you use color to convey information in the interface, you should also use clear, secondary cues to convey the information to those who cannot see the colors presented.
Levels of consistency	Make sure consistency is acquired on all levels: platform, suite of products, overall look & feel of a single app, icons, symbols, etc.
Consistency with user expectation	The most important consistency is consistency with user expectations
Defaults	Defaults within fields should be easy to "blow away."
Discoverability	If the user cannot find it, it does not exist.
Fitts's Law	The time to acquire a target is a function of the distance to and size of the target.
Human-Interface Objects	Human-interface objects have standard resulting behaviors and should be understandable, self-consistent, and stable.
Latency Reduction	Keep users informed when they face delay.
Readability	Favor particularly large characters for the actual data you intend to display as opposed to labels and instructions.
Simplicity	Avoid the "Illusion of Simplicity": simplicity is achieved by simplifying things, not hiding things.

All these principles are unique in meaning; the solutions in QEA 2.0 do, however, often relate to multiple principles. As such, the principles will not be discussed individually. In this section, we shall confine ourselves to three specific examples to clarify the types of changes made. In general, all changes were made to increase the level of consistency of the QEA, to increase the visibility of important elements, and to stay close to user expectations and user experiences with other software.

The first change is concerns the indication of the number of sources of evidence connected to an inference within an interpretive argument. In QEA 1.0, this indication differed for a number of sources of evidence connected to domains

and inferences. For consistency reasons, in QEA 2.0, the indication looks the same and only appears when one or more sources of evidence are added. This can be seen in Figure 7.8 where the evidence tab is displayed and where sources of evidence are connected to several domains and inferences.
A second example is the navigation within the different modules. Once a user chooses a module, several screens can be opened. Within the evaluation module, for example, users can either open the evaluation systems or the results page. The navigation through these screens in QEA 1.0 was always through the Home screen of the module. In QEA 2.0, back buttons were added to make navigation more similar to, for example, internet browsers. By choosing design elements based on broadly available and well-known software, the expectation is that usability will become more intuitive.
The third and last example has to do with the colors in the evaluation module. Users were confused by the number of colors and the different meanings of the colors in QEA 1.0. Therefore, the color scheme of QEA 2.0 was adjusted to be more consistent throughout the program. This specific coloring is presented in Figure 7.9, which displays a screenshot of the evaluation module with the new color scheme.

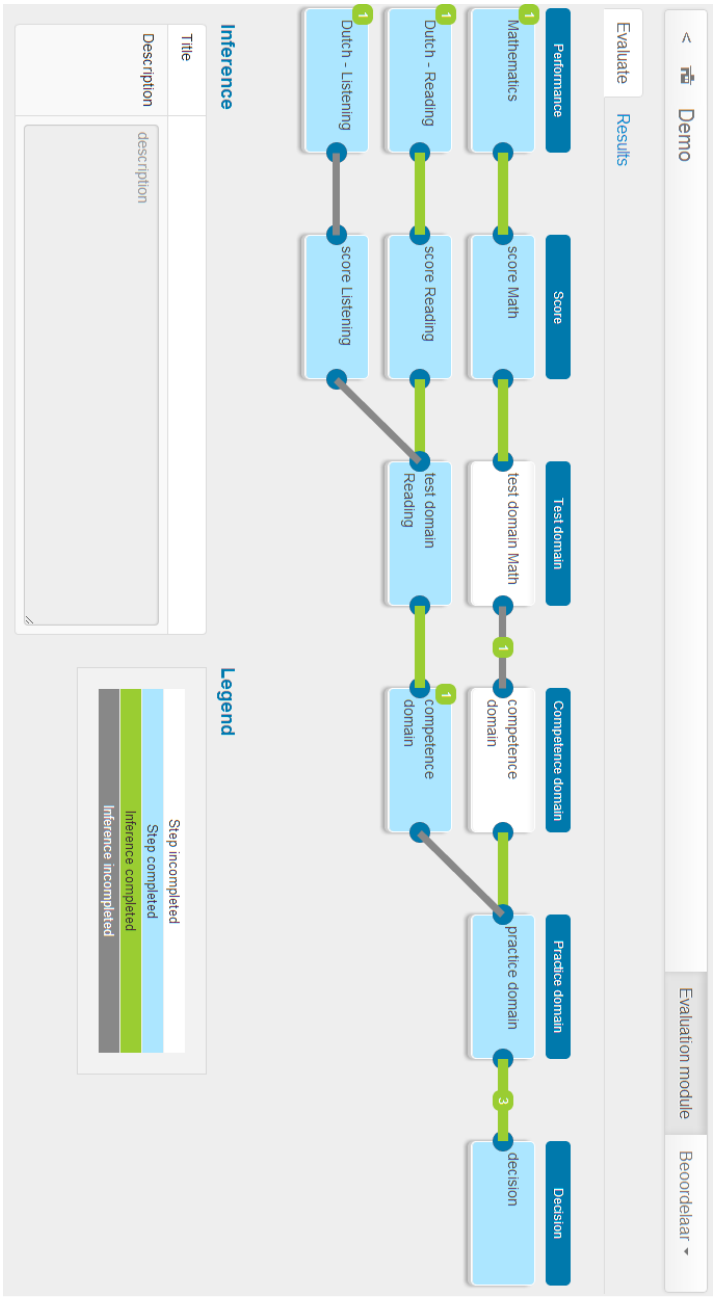


Figure 7.9: Evaluation tab in QEA 2.0 – used to select domains and inferences for evaluation

Study 2

The purpose of Study 2 was to evaluate QEA 2.0 and to get an indication of whether the changes represented an improvement over QEA 1.0. To do so, Study 2 also included a focus group, which was asked about the software through questionnaires. Two research questions were posed for this study.

2.1 Are stakeholders agreeing with the theoretical framework and underlying design choices made while implementing the argument-based approach to quality into QEA 2.0? Does this differ between the two studies due to the different designs?
2.2 Does the interface of QEA 2.0 enhance the clarity and usability of the software over the interface of QEA 1.0?

Methods Study 2

Procedure

The focus group was organized in 2014 during the Innovations in Testing conference by the Association of Test Publishers (ATP) in Scottsdale, AZ (Wools, 2014). Conference participants attended a 60-minute workshop in which a demonstration of the software was alternated with filling out questionnaires about QEA 2.0. During the session, information was given on particular parts of QEA 2.0 in two rounds. A questionnaire was completed in both rounds. During the rounds, the aim was to give just enough information for participants to be able to fill the questionnaires, but not too much so that they could respond spontaneously.

Materials

Like Study 1, in this workshop, two different types of questionnaires were used relating either to the theoretical framework (T-Questionnaires) or the interface and usability (I-questionnaires) of QEA 2.0. Due to time constraints, each questionnaire consisted of two parts instead of three, which was the case in Study 1. Because of the adjustments in the software and the availability of only two parts within the questionnaires, it was not possible to include all questions from Study 1. Therefore, a selection of items from the questionnaires used in Study 1 was selected for Study 2.
The T-Questionnaires addressed questions relating to the theoretical model incorporated in the software. The selected questions were the least likely to be

impacted by the changed interface, for example, questions relating to the prioritization of the initial design principles or questions requiring that participants write down examples of evidence.
The questions on both questionnaires were coded as correct or incorrect when applicable or, when necessary, into more comprehensive categories.
The I-Questionnaires focused on questions aimed at determining whether the adjustments of the interface could be considered an improvement. Therefore, questions concerned with aspects that were changed between versions 1.0 and 2.0 were selected for the questionnaire in Study 2. The questions focused on the indication of the number of sources of evidence connected to a component of an argument, the buttons used to build arguments, and the new color scheme.

Participants

The number of respondents in the two rounds differs since the questionnaires were distributed upon request (Theoretical, Interface, or both), and some respondents did not have enough time to complete every questionnaire. Table 7.11 displays the number of respondents for each round and questionnaire. In total, 12 participants handed in one or more questionnaires.

Table 7.11: Number of respondents per round and questionnaire

Round	Questionnaire T	Questionnaire I
1	9	12
2	8	12

Results Study 2

In the description of the results of Study 2, the same categories are distinguished as in Study 1. Within the description of the results, the results of Study 2 are presented as well as a comparison with the results of Study 1. For clarity purposes, therefore, some of the earlier reported results are repeated.

Theoretical Framework Questionnaires

This section addresses the results of the questionnaire in relation to the theoretical framework underlying the QEA. All results are presented to answer the first research question: are stakeholders agreeing with the theoretical framework and underlying design choices made while implementing the argument-based approach to quality in QEA 2.0? Does this differ between the two studies?

1. Sources of evidence.

When participants were asked for examples of evidence for a given Toulmin argument that consisted of two rebuttals, a warrant, and a backing, the same sources of evidence were mentioned as in Study 1. Only one additional example was mentioned to support the warrant.
Table 7.12 displays the sources of evidence mentioned in both studies. For both studies, the frequency with which the examples were mentioned is given. In Study 1, only five respondents answered these questions. In Study 2, seven to nine respondents answered them.

Table 7.12: Examples of evidence given by participants in Study 1 and Study 2

Claim within scoring inference	Examples of evidence	Study 1	Study 2
Rebuttal 1: Raters are not agreeing on the scores	• rating form from two raters	3	1
	• inter-rater reliability study and results	2	8
	• notes on discussion session among raters	1	0
Warrant: Scores are given by trained raters	• description of training program	3	1
	• report on their working procedure	1	2
	• raters are certified, present proof of competence or experience	0	6
	• in-depth qualitative research on how raters use the scale	1	0
Rebuttal 2: Raters do not follow the rating scheme	• examples and number of occurrences of deviations from rating scheme	1	6
	• interview raters on their use of the rating scheme	1	1
Backing: Raters use a rating scheme	• the rating scheme that is used	3	6
	• an archive of all rating schemes	2	1

2. Evaluation of evidence

As in Study 1, the participants in Study 2 were presented with several claims within a Toulmin argument and the accompanying evidence. Respondents were asked to evaluate the sources of evidence and decide whether they accepted or rejected the claim or whether they thought more evidence was needed to decide. As in Study 1, the results varied for different respondents. All judgments are presented in Table 7.13.

Table 7.13: Evaluation of evidence in Study 2

Participant ID	Warrant: Cut scores & norms are available	Rebuttal: Procedures to set standards are not carefully performed	Backing: Criterion-referenced norms based on content descriptions
27	Accepted	Accepted	Accepted
28	Accepted	Unclear	Unclear
29	Accepted	Unclear	Rejected
31	Accepted	Rejected	Accepted
32	Accepted	Unclear	Unclear
33	Accepted	Accepted	Rejected
36	Unclear	Rejected	Rejected
37	Accepted	Rejected	Accepted

3. Design principles

In the current study, the questions relating to the design principles were repeated from Study 1. As expected, the results obtained from these questions in Study 2 did not differ from those obtained in Study 1 (see Table 7.4 for Study 1 results). The respondents from Study 2 also regarded the first three design principles as most important, with the first principle seen as the most important category.

Interface and Usability Questionnaires

The results in this section aim to answer the second research question: does the interface of QEA 2.0 enhance the clarity and usability of the software over the interface of QEA 1.0?

4. Clarity of the modules

In QEA 2.0, a distinction is made between a test developer's area, with a design module and an evidence module, and an evaluation area, formerly referred to as an evaluation module. In Study 1, respondents were able to identify the purpose of the evaluation module once this was posed as a multiple choice question. Therefore, the interface was only tested for the design and evidence modules, both incorporated in the test development area in QEA 2.0. Respondents were asked to choose the one purpose out of three that described the goal of the modules. In Study 2, seven out of 12 respondents chose the right purpose for the design module, and nine out of 12 chose the right description for the evidence module. In comparison, in Study 1 (Table 7.6) the results were two correct out of 15 participants for the design module and nine correct out of

15 participants for the evidence module. Thus, in Study 2, the question relating to the design module was answered correctly more often than in Study 1. For the evidence module, this was the other way around.

5. Building arguments

Table 7.14: Scored responses on interface questions in Study 1 and Study 2.

	Study 1 (QEA 1.0)		**Study 2 (QEA 2.0)**	
Question:	N	Correct	N	Correct
Q1: What do you think will happen if you press the… Study 1: button "Add test domain?" Study 2: **+** button next to Test domain	13	12	12	12
Q2. You would like to add an additional rebuttal … What do you need to do to achieve this?*	10	6	12	7
Q3: What do you think is necessary to build complete the argument?*	11	9	12	7

*Questions were identical in Study 1 and Study 2, however, the interface changed.

As in Study 1, participants were asked to answer questions that could be interpreted as indicators of the usability of the software. When two questions were posed, the one receiving a higher frequency of correct answers was interpreted as easier. It was hypothesized that the new interface would be easier to use, and therefore, the questions posed in relation to this interface should become easier.

Table 7.14 displays the number of respondents who answered the questions, the number of correct responses, and the number of incorrect responses. In this table, the first question is divided into two questions: one posed in Study 1 and one posed in Study 2. However, this only concerned a change in wording to match the new interface. The functionality to which these questions referred remained the same. Noteworthy, the first two questions regarding QEA 2.0 in Study 2 were more difficult than the questions on QEA 1.0. This could indicate that the interface did not add to the usability but actually made it more complex. The last question did not differ in either version of the QEA.

To get an understanding of the added complexity in the new version of QEA 2.0, the responses of the participants are evaluated more thoroughly. The first question concerned adding a test domain by clicking the "+" button. Participants who answered this question incorrectly thought that clicking "+"

would unveil underlying elements that were not yet displayed. The incorrect answers given for the questions on "adding a rebuttal" were either respondents who did not know the answer or participants who described the process theoretically. An example of an answer in the latter category was: "your addition should be nested within a relation to a rebuttal. You should formulate an R2 (probably an R3) with a B" (respondent 34). Finally, the third question about completing an argument was answered incorrectly by respondents who described the process of building arguments in a general way. For example: "plan, design, build, analyze" (respondent 31).

6. Indication of sources of evidence

One aspect of particular interest was the number of sources of evidence connected to an inference or domain and the indication showing this number. In Study 1, this indication was not recognized as a representation of the number of sources of evidence connected to an element. Therefore, this indication was adjusted in QEA 2.0, and the questions relating to this number were repeated. In Study 2, seven out of eleven respondents were able to correctly identify the function of the indication of sources of evidence. Furthermore, all seven respondents recognized that the meaning of both the number displayed on the top of an inference as well as the number displayed on the top of a domain held the same interpretation.

7. Evaluation module

The final part of the I-Questionnaire focused on the color scheme used in the evaluation module. The results of Study 1 indicated that respondents had difficulty recognizing the meaning of the colors used, even when a legend was provided. The color scheme was adjusted for QEA 2.0. Three questions from Study 1 were repeated in Study 2. In Study 2, two questions were answered correctly by eight respondents, and one question was answered correctly by six respondents out of twelve. This means that in Study 2, the majority of the respondents were able to correctly describe the meaning of the colors.

Conclusion and Discussion

This paper discussed the evaluation of the quality of tests and assessments. The definition of quality used here is the degree to which an assessment instrument is useful for its intended purpose. To evaluate quality, which is dependent on

the intended purpose of assessments, an argument-based approach to quality was used. This approach is an extension of the argument-based approach to validation proposed by Kane (2006). One of the problems related to this approach is that it is difficult for practitioners to use the principles and framework proposed in the theoretical work. To support practitioners in their validation efforts and the evaluation of assessment quality, this paper presented software that aims to guide users through this process.

The software – Quality Evaluation Application (QEA) – aims to help users in building quality arguments, supports the storage and classification of sources of evidence, and structures the collected sources of evidence. Due to this structured approach, users can build a quality portfolio that can be evaluated as a whole, which might enhance the comparability of auditing efforts by different auditors.

The software was developed in a design-based research project (McKenney & Reeves, 2012). Following the rationale of design-based research, design principles were formulated based on a theoretical framework. The four design-principles in the current project were translated into a prototype of the software: QEA 1.0, and a study (Study 1) was conducted to evaluate the prototype. As part of this study, a focus group was organized to answer three research questions on three out of four design principles. The participants of the focus group filled out several questionnaires on the software prototype. The results of Study 1 were subsequently incorporated in a redesign of the software. During this phase, the original design principles were used to prioritize the different suggestions for improving the first prototype. An adjusted version (QEA 2.0) was developed in line with these suggestions. In the new version of the software, the structure was adjusted, and the interface design was changed so that the program was more clear. The QEA 2.0 version was evaluated during Study 2. A focus group was also held for this second study. In the questionnaires completed during this focus group, questions from Study 1 were repeated in relation to the new interface. This enabled comparisons between the usability of the two versions of the software.

Conclusion

In this section, conclusions are drawn on the extent to which the developed software fits the initial design principles. Therefore, the four principles are repeated, and conclusions are grouped according to them.

1. The system can be used during several stages of the test development process: it is not only suitable to guide test development but can also be used as an instrument for internal or external audits.

In this study, users indicated on several occasions that they thought the software would be most useful during the test development stage. Building the arguments would force test developers and stakeholders to discuss design choices in detail. Furthermore, the underlying claims and assumptions implied by certain design choices would be elucidated. When design choices are discussed, debated, and challenged extensively during the development process, this could presumably lead to higher quality assessments. However, facilitating discussions related to test development was not a primary purpose of the software.

The main purpose of the software is to facilitate the evaluation of assessment quality. One aspect of this functionality is that users need to be capable of thinking about sources of evidence that they can present to support claims and assumptions made within the quality argument. During both Study 1 and Study 2, respondents were asked to name sources of evidence that they would provide to support claims in the questionnaire. The results show that most respondents were capable of listing examples of sources of evidence that they would present. Furthermore, although the examples differed somewhat, the variability was not very large. Therefore, we can conclude that potential users are able to adequately list sources of evidence and that they are capable of using the theoretical framework to build quality portfolios.

Whether the software can also be used for internal and external audits is still unclear. Although during both evaluation studies, questions were posed in relation to the evaluation of presented evidence, it remains unclear whether the judgments are more comparable and whether auditors are able to evaluate a complete assessment. To evaluate this particular use of the software, it is necessary to prepare a quality argument and accompanying evidence that can be evaluated by auditors in an experimental setting. However, to do so, the software needs to be sufficiently usable for users to work with it without support. Therefore, usability and interface are a prerequisite for comparison studies.

2. The system defines quality as the extent to which something is useful for its purpose and, therefore, incorporates an argument-based approach to quality.

During the evaluation studies, respondents were asked to answer questions relating to specific design choices regarding the implementation of the argument-based approach, for example, the restriction of the software to add multiple decisions in one argument. Respondents agreed with such uses. They also indicated that they thought the software and the underlying definition of quality were useful. The issues mentioned during the focus groups were in relation to the functionality of the system or aspects that were unclear in the argument-based approach. The latter could also be interpreted as an indication of the necessity for guidance in using the argument-based approach.
This project also evaluated the design principles. Respondents were asked to prioritize these principles and to add any missing aspects. Only a few respondents did actually suggest additional principles, but the decision was made that these additional principles were already subsumed within the existing four principles.

3. The auditing process is simplified, and auditors' judgments should become more comparable.

In this project, this particular design principle was interpreted more broadly. The focus was extended from the auditing process to usability and full functionality. Thus, the focus was also on the usability of the area in the software where test developers build their quality arguments. It was thought that the process is only simplified once the software is easy to use and works intuitively. Furthermore, it was preferable that potential users could use and understand the interface without additional training in the use of the software. This is because of the extent of the actual theoretical knowledge necessary to build these kinds of validity or quality arguments. Therefore, is was deemed ineffective to extend the burden of training and explanation with complex software.
Two versions of an interface design were evaluated. The results indicate that the newer version is somewhat more useful because some misconceptions from Study 1 were not encountered in Study 2, for example, in relation to the indication of the number of sources of evidence classified within the validity argument. However, not all changes led to a better understanding of the software. This could be concluded from the fact that not all repeated questions were more frequently answered correctly in Study 2. To get a better understanding of these aspects, the written answers were evaluated in greater detail. From this evaluation, it seems that the persistent misunderstandings concerning functionality might be easy to solve. It could also be the case that

these misunderstanding disappear once users can actually work with the software. This latter hypothesis is discussed more extensively in the discussion part of this section.

4. The system should include other evaluation systems to prevent it from being just another evaluation system.

The current software incorporates two evaluation systems: the ABA and the Dutch COTAN system. It is possible to add more systems, but this was not prioritized over building and evaluating the usability of the software. Furthermore, the current evaluation systems and their implementation need to be extensively evaluated before adding further systems.

Discussion

In this research project, we developed software that can guide and support evaluation practice in evaluating the quality of tests and assessments. We concluded that the current software is regarded useful and that there is a consensus on the implementation of the theoretical framework. Despite these positive outcomes for the software, some aspects relating to the research project require further discussion.

First of all, the process of test development and the evaluation of this development efforts are still complex. This means that building quality arguments, making design choices that fit the intended use of assessments, and evaluating whether test developers succeed in these efforts are tasks that require specific expertise. The competence and expertise necessary to perform these tasks are very diverse and are often divided over multiple persons. Therefore, constructing assessments, planning the validation efforts concerning these assessments, and evaluating them need to be a team effort. The current software does incorporate the possibility of building quality arguments with multiple users. The evaluation module is unfortunately not yet suited to divide aspects for evaluation between auditors with different kinds of expertise.

The second aspect regarding the assessment evaluation is our understanding of the measurable construct "quality." When assessments are evaluated, this could be interpreted as assessing the quality of assessments. In educational measurement, this would often mean that we are interested in an equivalent of a "true score" while in quality evaluation, this would mean that we are interested in identifying the "true quality" of a test. Consequently, this would mean that all variability in the ratings of the quality of a test should be regarded

as construct irrelevant variance. However, once we adopt a definition of quality that allows for different auditor perspectives, for example, because they have different values, we should also allow for variability in their judgments. When auditors do not agree on the status of evidence for assessment quality, in this latter view, this could be regarded as construct relevant variance. This would be relevant since the actual difference in judgments matters in getting a full understanding of the extent to which an assessment is useful for an intended purpose, given different perspectives.

This latter view on test quality does however cause problems once test developers need to demonstrate that their test is good enough for a particular purpose. Once this is the target of the evaluation, an instrument that incorporates norms should be used, such as the COTAN system (Evers, et al., 2010) or the RCEC evaluation system (RCEC, 2015). These norms should guide auditors in their judgment and would make the individual values of auditors obsolete. This would also mean that auditing systems, which incorporate norms, should evaluate the inter-rater-reliability of auditors and that differences between ratings be interpreted as construct irrelevant variance.

Alongside aspects relating to the evaluation of assessments, some points regarding design research, and this project in particular, also need mentioning.

In this particular project, questionnaires were used to evaluate the interface of the software. Although it is useful to investigate whether potential users can explain or describe the functionality they expect in a particular button or area of the software, it is not a very authentic measure. This means that conclusions regarding the actual usability of software need to be drawn with caution. This is especially the case for conclusions about aspects of the software that are misinterpreted or that might not work. All these aspects might, for example, be solved once a potential user gets an opportunity to work with the software and to discover through trial-and-error the functionality of buttons. Also, in the project described in this paper, it might be the case that the misunderstandings that users have are exaggerated because participants could not work with the software themselves. It is important that follow-up studies focus on evaluating the learning curve that users need to go through before they can use the software.

Another point that could have influenced the conclusion of the evaluation studies were the different circumstances of the focus groups. The interaction during the workshops, but also the demonstration of the software, differed, and therefore, the two groups of participants did not have the same amount of knowledge before answering the questionnaires. This latter point was assumed

when the answers of both groups were compared. Furthermore, the focus groups were organized during two different conferences. The participants of these conferences differed in background knowledge and interest. Therefore, the results of the comparison should be regarded as informative but should not be interpreted as decisive evidence.

As mentioned earlier, this project was performed on the basis of the principles of design-based research. One of the main aspects of this research approach is its iterative nature (Plomp, 2007). During the course of this project, there were several points during the formal cycles where feedback was provided and when the necessary software was immediately adjusted. This process, whereby small improvements or informal moments of feedback are repeated in quick succession, does not fit the practice of reporting in scientific discourse. Therefore, this paper reports on the formal cycles, which were planned and followed, and all small changes were taken together in the overarching steps described. Unfortunately, this leaves us with a description that does not cover the dynamic interaction and frequency of feedback that is usually obtained in these kinds of research projects.

This paper provides a description of the development of the QEA, and the next phase in this design-based research project will comprise of the implementation of the QEA into existing practices. This could be seen as the next step in developing the software and will undoubtedly lead to more changes. It is important to weigh these new suggestions for improvement against the existing design principles and the existing knowledge built around this particular software.

As a closing remark, we would like to emphasize that developing a software within a design-based research project is very much like developing and validating assessments. The main activities are to keep challenging design choices, to weigh them against initial design principles, and, most importantly, to keep collecting and interpreting evidence that supports or rejects claims and assumptions that are implied in design choices.

References

AEA-Europe. (2012). *European framework of standards for educational assessment.* Retrieved from http://www.aea-europe.net/index.php/professional-development/standards-for-educational-assessment

American Educational Research Association (AERA). American Psychological Association (APA), National Council on Measurement in Education (NCME).

(1999), *Standards for educational and psychological testing.* Washington: American Psychological Association.

Bartram, D. (2001). The development of international guidelines on test use: The international test commission project. *International Journal of Testing, 1*(1), 33–53.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP.

McKenney, S., & Reeves, T. (2012). *Conducting educational design research: What, why and how*. London: Routledge.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Kane, M. T. (2013), Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice, 27*(3), 32–42.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Charlotte, NC: Information Age Publishers.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, 99–104. doi: 10.1111/jedm.12005

Tognazzini, B. (2014). First principles of interaction design (Revised & Expanded). Retrieved from: http://asktog.com/atc/principles-of-interaction-design/

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The uses of argument* (Updated Edition). Cambridge: Cambridge University Press.

Plomp, T. (2007). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 9–35). Enschede, Nederland: SLO.

RCEC. (2015). *Het RCEC Beoordelingssysteem.* Enschede: RCEC.

Wools, S. (2012). Towards a comprehensive evaluation system for the quality of tests and assessments. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 95-106). Enschede: RCEC.

Wools, S. (2013). *Quality of assessments and assessment programs: Demonstration of an online evaluation system.* Workshop presented at EARLI, München.

Wools, S. (2014). *Mirror mirror on the wall, who has the best assessment of them all? Presenting a new approach for the online evaluation of assessments.* Workshop presented at ATP Innovations in Testing 2014, Phoenix, AZ.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63–82.

Wools, S., Den Otter, D. & Eggen, T. (submitted). A systematic literature review of validation studies on assessment.

Wools, S., Eggen, T., & Béguin, A. (submitted). Constructing validity arguments for test combinations.

Appendix I

Questionnaires used during Study 1

T-Questionnaire A – Quality Evaluation Application & Design Specifications	
1a	Prioritize these four design specifications by using 1, 2, 3, and 4, where 1 means the most important specification, and 4 means the least important specification.
1b	Do you feel that a specification is missing? If yes, please specify which one.
T-Questionnaire B – Design Module & Evidence Module	
2a	Do you agree of with the restriction of one practice domain? Can you think of a scenario where two practice domains are possible?
2b	Please write down an example where two scores lead to one test domain.
3a	From the perspective of quality evaluation, do you think it is necessary to add evidence for every element of the argument? Please elaborate on you answer with an example.
4	Can you think of sources of evidence that would support or reject the claims made within this argument?
4a	Rebuttal 1 – Raters are not agreeing on the scores
4b	Warrant – Scores are given by trained raters
4c	Rebuttal 2 – Raters do not follow the rating scheme
4d	Backing – Raters use a rating scheme
T-Questionnaire C – Evaluation Module	
5	Please consider the sources of evidence and evaluate the claims within this particular Toulmin argument. Decide for every element whether you think it is Accepted, Rejected, or Unclear.
5a	Warrant – Cut scores & norms are available
5b	Rebuttal – Procedures to set standards are not carefully performed
5c	Backing – Criterion-referenced norms based on content descriptions
6	Are there any other suggestions or comments that you would like to share?

I-Questionnaire A – Quality Evaluation Application & Design Principles	
1	Look at Figure 1 in Appendix I. You'll see a "module selection screen." Please write down what you think is the main purpose of the modules that can be selected.
1a	Design module. In this module, a user can …
1b	Evidence module. In this module, a user can …
1c	Evaluation module. In this module, a user can …
2a	What activity do you think is supposed to be performed in the Design module? A. Construct items B. Develop a test C. Prepare an evaluation of a test

	I-Questionnaire A – Quality Evaluation Application & Design Principles
2b	What activity do you think is supposed to be performed in the Evidence module? A. Gather evidence to show the competence of students B. Performing analyses – the results can be used as evidence of quality C. Store and classify evidence of test quality
2c	What activity do you think is supposed to be performed in the Design module? A. Evaluate assessment quality B. Evaluate students' abilities or competences C. Evaluate the functionality of the software
	I-Questionnaire B – Design Module & Evidence Module
3	What do you think will happen if you press the button "Add test domain"?
4	Look at the Toulmin model at the bottom of Figure 2. You would like to add an additional rebuttal for the warrant "scores are given by trained raters." What do you need to do to achieve this?
5	What do you think is necessary to build a complete argument in Figure 3? (It is not necessary to fill all the steps)
6	Take a look at the numbers shown in the argument graph in Figure 4.
6a	What do you think is the meaning of the numbers in green?
6b	What do you think is the meaning of the number on the pink line?
7	Figure 4 (Appendix II) shows an argument for an assessment program in mathematics and reading. As a source of evidence for the Test domain, someone added a Test matrix.xlsx with a test matrix for math. Unfortunately, this matrix for math was wrongly classified under the test domain 'Reading'. Can you indicate what needs to be done to correct this mistake?
	I-Questionnaire C – Evaluation Module
8	Indicate all the elements in the picture below that need to be evaluated for a complete evaluation of this assessment. You can mark the elements or lines that need to be evaluated by writing an "x" in every element.
9	Suppose that, in your opinion, the claim that is made in the rebuttal (see below) is proven to be right. This means that the test developer presented evidence that convinced you that "the procedures to set standd are not carefully performed." In which picture does the blue square indicate this opinion?
10	In Figure 6 (Appendix III), the domains, and the lines between them, have different colors. In the QEA, a legend is added. By means of this legend, can you indicate what you think a color means?
10a	A blue square for a domain means:
10b	A white square for a domain means:
10c	A green line between domains means:
10d	A pink line between domains means:
10e	A blue line between domains means:
11	In Figure 7 in Appendix III, the results of the evaluation are shown. There are

	I-Questionnaire A – Quality Evaluation Application & Design Principles
	three options for the results of the Toulmin argument: an argument is rejected, accepted, or unclear.
11a	Red: rejected/accepted/unclear
11b	Blue: rejected/accepted/unclear
11c	Green: rejected/accepted/unclear
12	Are there any other suggestions or comments that you would like to share?

Appendix II

Questionnaires used for Study 2

	T-Questionnaire A – Quality Evaluation Application
1a	Prioritize these four design specifications by using 1, 2, 3, and 4, where 1 means the most important specification, and 4 means the least important specification.
1b	Do you feel a specification is missing? If yes, please specify which one.
4	Can you think of sources of evidence that would support or reject the claims made within this argument?
4a	Rebuttal 1 – Raters are not agreeing on the scores
4b	Warrant – Scores are given by trained raters
4c	Rebuttal 2 – Raters do not follow the rating scheme
4d	Backing – Raters use a rating scheme
	T-Questionnaire B – Evaluation Module
5	Please consider the sources of evidence, and evaluate the claims within this particular Toulmin argument. Decide for every element whether you think it is Accepted, Rejected, or Unclear.
5a	Warrant – Cut scores & norms are available
5b	Rebuttal – Procedures to set standards are not carefully performed
5c	Backing – Criterion-referenced norms based on content descriptions
6	Are there any other suggestions or comments that you would like to share?
	I-Questionnaire A – Quality Evaluation Application
1a	What activity do you think is supposed to be performed in the Design module? A. Construct items B. Develop a test C. Prepare an evaluation of a test
1b	What activity do you think is supposed to be performed in the Evidence module? A. Gather evidence to show the competence of students B. Performing analyses – the results can be used as evidence of quality C. Store and classify evidence of test quality
2	What do you think will happen if you press the **+** button next to Test Domain
3	Look at the Toulmin model at the bottom of Figure 2. You would like to add an additional rebuttal for the warrant "scores are given by trained raters." What do you need to do to achieve this?
4	What do you think is necessary to build a complete argument in Figure 3? (It is not necessary to fill all the steps)
5	Take a look at the numbers shown in the argument graph in Figure 4. What do you think is the meaning of the numbers in green?
5a	Circle (inference)
5b	Square (domain)

	I-Questionnaire B – Evaluation Module
6	Indicate all the elements in the picture below that need to be evaluated for a complete evaluation of this assessment. You can mark the elements or lines that need to be evaluated by writing an "x" in every element.
7	Suppose that, in your opinion, the claim made in the rebuttal (see below) is proven to be right. This means that the test developer presented evidence that convinced you that "the procedures to set standard are not carefully performed." In which picture does the blue square indicate this opinion?
8	In Figure 6 (Appendix III), the domains, and the lines between them, have different colors. In the QEA, a legend is added. By means of this legend, can you indicate what you think a color means?
8a	A blue square for a domain means:
8b	A white square for a domain means:
8c	A green line between domains means:
8d	A pink line between domains means: (invalid question)*
8e	A blue line between domains means: (invalid question)*
9	Are there any other suggestions or comments that you would like to share?

*Due to an error in the construction of the questionnaire, two invalid questions were posed. During the session, participants were instructed not to answer these questions. The responses to these two questions were not taken into account during the analyses.

Chapter 8

Final considerations

The main idea presented in this dissertation is that validation research and quality evaluation of educational assessments are closely related. The usability of the argument-based approach to validation for quality evaluation purposes was shown in a design-based research project. In this study, a theoretical framework on validation theory was extended and used as the basis for a quality evaluation system for educational assessment. This epilogue raises some discussion points relating to the research project as a whole.

Design-based research

Design-based research is an applied research method used in this study. Its aim is to systematically develop a product or intervention with a strong theoretical rationale (McKenny & Reeves, 2012; Plomp, 2007). This kind of research is practical in nature and enables opportunities to employ research findings in products that are accessible to a broader public. One could argue that this answers to a call from society that research should be more focused on solving real life problems and that the value of research should be clear from the start. However, we should also consider that these kinds of research projects are only possible when fundamental knowledge is available. One could argue that it is a second step: first, there is fundamental research, which is then followed by design-based research as an implementation of what we know.

We could distinguish methods that provide us with fundamental knowledge from applied research methods, such as design-based research. These research approaches differ, for example, in the extent to which researchers are flexible during the course of a study. Theoretical scientific research is systematic, reproducible, and according to predefined research designs. Design-based research is also systematic but allows for adaptations to hypotheses during the course of a project. In a way, these design-based research methods are very much like the lean start-up methodology (Ries, 2011): while working on a project, you should always be able to pivot and change directions quite easily when you notice that something does not work.

Although design-based research is a necessary addition to fundamental or theoretical research, it is still not valued as such. Arguably, being able to make research findings accessible in products that can be of use to society should be regarded as highly as fundamental research.

Unfortunately, many journals are more likely to publish theoretical or fundamental research articles. One of the reasons that journal editors are not inclined to publish design-based research is that these articles are structured

very differently. Due to the multiple iterations and the dynamic character of design-based research, a traditional article structure is not suitable. I feel that being able to share our knowledge from design-based research should be made possible in all journals and not just methods-specific ones. Therefore, we need to make an effort to clearly present our research, but editors also need to be flexible in terms of the structure of articles that they would like to publish. Furthermore, as a community, we should think of new ways of presenting our knowledge to colleagues, possible users, and society.

Developing educational assessments of good quality

Quality evaluation is an activity that is usually performed at the end of a development cycle. We could compare this with an assessment or exam. At the end of a learning cycle, we would like to know whether we succeeded in achieving the learning objectives. This also means that evaluation in itself cannot enhance quality. When one wants to increase test scores, one should improve student learning. This is also true for assessment quality. If we would like to improve the quality of our assessment, we should improve its making.
Unfortunately, this is significantly easier said than done. Increasingly, the availability of ICT, complex psychometrics, and the growing demands for efficiency in testing make constructing good measuring instruments a challenge, one we can only solve when we approach this as a team effort with content experts, psychometricians, potential users, and IT-specialists working together. This could only work when everyone knows what s/he needs to do, what s/he is responsible for, as well as when everyone has the end result in mind. Developing assessments does not involve gluing different pieces together into one test; it involves making design choices that are coherent with one another. Moreover, when choices need to be reconsidered due to practical challenges, all other design choices should also be validated. One way to do this is by keeping track of the development process. I feel that the quality evaluation application (QEA), which was developed during our research project, might be of help here: a formative instrument that supports test construction and guides research efforts performed during the construction process.

Sound assessment use

From an assessment construction point of view, developing an assessment that is suitable for an intended interpretation and use is within scope. However, whether this instrument is actually used for this intended interpretation and use is not something a test publisher can control.
There are indeed some aspects that can be seen as part of test construction, which would help users in putting assessment results to good and appropriate use. The most important component is score reporting. When it comes to the intended interpretation and use of assessment scores, one of the most important aspects is whether we can appropriately present these results. When we neglect this stage of our assessment design, all efforts put into delivering valid test results are lost in translation. Therefore, score reports should always be rigorously evaluated (Van der Kleij & Eggen, 2013). Some important questions include: are users able to understand the reports, and are they able to draw valid conclusions based on these reports? Of course, this is not only a matter of clear reporting; test users, such as teachers, also need sufficient assessment literacy (Fullan & Watson, 2000) to understand the possibilities as well as the limitations of test scores. In the Dutch context of higher education, several recent initiatives aim to enhance assessment literacy in teachers and policymakers (Expertgroep BKE/SKE, 2012; Van Berkel, Sluijsmans, & Joosten-ten Brinke, 2015). In the end, however, the quality of the assessment, the clarity of the report, and the assessment literacy of teachers are probably all factors of importance when it comes to sound assessment use.

Validity versus quality?

As a final point, throughout this dissertation, validity is described as distinct from quality. However, when we compare the two definitions used in this thesis, it becomes clear that this distinction is questionable to say the least.

Validity:

> 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. (Chapter 3, p. 46)

Quality:

> In the new evaluation system presented here, quality is defined as the degree to which an assessment instrument is appropriate for its intended purpose. …The central idea behind this system is for it to be used to build an argument that helps test developers demonstrate that a test or assessment is sufficiently useful for its intended purpose. To build this argument, evidence is needed to convince all stakeholders of a test's usability. (Chapter 7, p. 152)

In comparing these two definitions, it becomes clear that during this research project, the definitions of validity and quality became very closely interrelated. Furthermore, in this dissertation, we have shown that the procedure to support validity can also be applied to quality. It makes me question the need to make a distinction between the two concepts when using them in practical contexts. The current debate (see, for instance, *Measurement: Interdisciplinary Research and Perspectives*, 2012; Clauser & Wells, 2013) on validity is considerably fierce, but in my opinion, it is a highly theoretical discussion. In practice, the difference between validity and quality might be of less relevance when we agree that in all instances, the appropriateness of an assessment to be used for a specific purpose must be demonstrated. Regardless of whether this is part of validation studies or quality evaluation, in the end, it is the same activity. I think that Newton and Shaw (2014) would classify me as a liberal or perhaps even a *hyper-liberal.* And I have to admit that during the course of this study, I did become an extremist on this part when I say: when it comes to quality, it's all about validity.

References

Expertgroep BKE/SKE (2013). *Verantwoord toetsen en beslissen in het hoger beroepsonderwijs. Een voorstel voor een programma van eisen voor een basis- en seniorkwalificatie examinering (BKE/SKE)*. Den Haag: Vereniging Hogescholen.

Fullan, M., & Watson, N. (2000). School-based management: Reconceptualizing to improve learning outcomes. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, *11*, 453–473.

Clauser, B. E., & Wells, C. S (Eds) (2013). Special Issue on Validity. *Journal of Educational Measurement*, *50*(1), 1–122.

McKenney, S., & Reeves, T. (2012). *Conducting educational design research: What, why and how*. London: Routledge.

Measurement: Interdisciplinary Research and Perspectives (2012). Issue 1 – Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), *1–122.*

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment.* London: Sage.

Plomp, T. (2007). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 9–35). Enschede, Nederland: SLO.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Charlotte, NC: Information Age Publishers.

Ries, E. (2011). *The lean startup.* USA: Random House USA Inc.

Van Berkel, A., Sluijsmans, D. M. A., & Joosten-ten Brinke, D. (2015). Kwaliteit van toetsbekwaamheid. In D. Sluijsmans, D. Joosten-ten Brinke, & T. Van Schilt-Mol (Eds.), *Kwaliteit van toetsing onder de loep* (pp. 133–149). Antwerpen; Garant.

Van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation*, 39, 144–152.

Summary

All About Validity

An evaluation system for the quality of educational assessment

At all levels of education, tests and assessments are used to gather information about students' skills and competences. This information can be used for decisions about groups of students or about individual students. When the stakes of these decisions are high, it becomes increasingly important that the assessments used are of good quality. In this dissertation, assessment quality is defined as the degree to which an assessment instrument is appropriate for its intended purpose. This definition entails a flexible view of quality: dependent on the intended interpretation and use of test scores, different quality criteria should be used for evaluation. This approach is also described as an argument-based approach to quality. It originates in validity theory, and our research project aimed to extend it to quality. This is done in a design-based research project whose aim was to develop an evaluation system, which incorporates an argument-based approach to quality, for the quality of educational assessments. In this design-based research project, a theoretical foundation is developed. From this theoretical foundation, design principles are formulated and then used to build a prototype of the evaluation system. This prototype is then evaluated and adjusted in an iterative process. This dissertation describes the theoretical foundation, the design principles, and several iterations in the development of the prototype.

Theoretical foundations

The original argument-based approach to validation consists of two stages. In the first stage, an interpretive argument is constructed. This argument aims to describe the intended interpretation and use of assessment scores. Furthermore, it seeks to elucidate the underlying inferences drawn in order to reason from the observed assessment performance to the intended decision. When all inferences are made explicit in the interpretive argument, a validity argument is written. This validity argument combines several sources of evidence that both support and reject claims made within the interpretive argument. This latter

argument provides us with a weighted conclusion on the validity of the assessment scores for a predefined interpretation and use.
To extend this approach to quality, a third stage is added. In this evaluation stage, three criteria are added to evaluate the interpretive argument, the evidence that is provided, and the validity argument. These extensions are described in depth in Chapter 2. The extended approach is illustrated with an existing assessment: the driver performance assessments. From this example, it is shown that the extension of the argument-based approach makes it possible to be used for evaluation purposes.
The argument-based approach to validation is described in the context of a single test or assessment. However, in educational assessments, tests or assessments are often combined to make one decision. These assessment programs are used, for example, for diploma decisions as well as to evaluate students' progress. Chapter 3 proposes an extension to the argument-based approach to the validation of multiple tests. This extension is illustrated with the validation of a competency assessment program (CAP) for social workers. This CAP is validated in collaboration with a quality manager of an educational program. The case study illustrates 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 and also guides more detailed validation efforts.
Following the two extensions of the argument-based approach to validation, Chapter 4 aims to put the approach to use in a complex situation. This chapter purports to show the advantages of the argument-based approach in gaining understanding about the quality of assessments. Furthermore, it shows that the argument-based approach facilitates researchers and policymakers in deciding whether particular design choices contribute to the quality of a decision made within an assessment programs. To do so, this chapter focuses on a new national assessment program in arithmetic in the Netherlands. It illustrates that the most important claims are related to the comparability of the individual components of the assessment program. Therefore, data is used to evaluate the level of comparability and to verify the claims being made within the program.
The argument-based approach to validity is well known in validity theory. However, it was unclear whether researchers adopted the approach. Therefore, our study looks at whether researchers performing validation studies differentiate the sources of evidence needed to support different intended uses of test scores. Chapter 5 describes a systematic literature review, which focuses

on the hypothesis that depending on the intended use of the tests, authors will present different sources of validity evidence. The literature review includes 178 articles on validation efforts in educational assessment. All sources of evidence presented by the researchers in these articles are classified within a theoretical model of the interpretive argument. We then analyze whether differences occurring in the presented evidence depend on the intended use of the test scores. The results show that this is the case for assessments constructed for selection purposes. The validity of these tests is more frequently supported with evidence relating to the possibility of accurately predicting future behavior. Tests with other intended uses do not differ in terms of the sources of evidence provided. The results also show that the majority of the articles present only one or two sources of evidence instead of evidence supporting a full validity argument.

Design principles

Based on the theoretical foundations, it was concluded that the argument-based approach to validity also seemed suitable for quality evaluation. It was also concluded that researchers did not adopt the rationale in their current practice. To support researchers and test developers in building arguments according to the argument-based approach to quality, design principles for an online evaluation system were formulated in Chapter 6. When researchers are supported by this software, they might be more inclined to differ in the sources of evidence needed to evidence the quality of different assessments.
The design principles in this chapter were derived from a comparison of currently available evaluation systems. The main difference for the new evaluation system is that it incorporates an argument-based approach to quality. Furthermore, it should be used during several stages of the test construction process. The third principle describes that the system should simplify the auditing process and that, therefore, auditors' judgments should become more comparable. Finally, the system should include other evaluation systems to prevent it from being just another addition to already existing systems.

Prototype

Chapter 7 presents an online evaluation system for the quality of tests: the Quality Evaluation Application (QEA). The QEA was developed in several iterations of a design-based research project, and the prototype of the software

was developed from the design principles. In this chapter, the development of the software is described as well as two evaluation studies on whether the initial design principles were met. The chapter concludes with the assertion that the software is promising to foster discussion during test construction. Furthermore, the interface seems sufficiently useful to initiate a new phase in the research project whereby the software can be implemented for use by practitioners.

Samenvatting

Alles Is Validiteit

Een beoordelingssysteem voor de kwaliteit van toetsen en assessments in het onderwijs

In alle onderwijsniveaus worden toetsen en assessments gebruikt om informatie te verzamelen over de vaardigheden en competenties van studenten. Deze informatie kan gebruikt worden om beslissingen over groepen of individuele studenten te nemen. Wanneer het belangrijke beslissingen betreft, is het van groot belang dat de toetsen die hiervoor gebruikt worden van goede kwaliteit zijn. In deze dissertatie wordt kwaliteit van toetsen gedefinieerd als de mate waarin een toets geschikt is voor het beoogde gebruik. Deze definitie past bij een flexibel beeld van kwaliteit: afhankelijk van de beoogde interpretatie en het beoogde gebruik van toetsscores worden verschillende criteria gebruikt om kwaliteit te evalueren. Dit concept kan ook wel omschreven worden als een argumentgerichte benadering van kwaliteit. Deze benadering komt oorspronkelijk uit theorieën over validiteit maar is in dit onderzoeksproject uitgebreid naar kwaliteit. Hiervoor is een ontwerpgericht onderzoek uitgevoerd met als doel een beoordelingssysteem voor de kwaliteit van onderwijskundige toetsen te ontwikkelen waarbij gebruik gemaakt werd van de argumentgerichte benadering van kwaliteit.
In het ontwerpgerichte onderzoek is eerst en theoretisch kader ontwikkeld. Vanuit dit kader zijn ontwerpprincipes geformuleerd en deze zijn vervolgens gebruikt om een prototype van het beoordelingssysteem te ontwikkelen. Het prototype werd daarna in een iteratief proces geëvalueerd en aangepast. In deze dissertatie worden het theoretisch kader, de ontwerpprincipes en twee iteraties van de ontwikkeling van het prototype beschreven.

Theoretisch kader

De oorspronkelijke argumentgerichte benadering van validiteit bestaat uit twee fasen. In de eerste fase wordt een interpretatief argument ontwikkeld. Dit argument is erop gericht de beoogde interpretaties en het gebruik van toetsscores te omschrijven. Daarnaast wordt in dit argument gepoogd om inferenties expliciet te maken die gedaan worden wanneer geredeneerd wordt

van een observeerbare toetsprestatie naar de beoogde beslissing. Wanneer alle inferenties expliciet gemaakt zijn in het interpretatieve argument wordt een validiteitsargument geschreven. In dit validiteitsargument worden verschillende bewijsbronnen gecombineerd waarmee de claims uit het interpretatieve argument ofwel ondersteund ofwel tegengesproken worden. Op basis van deze bewijsbronnen levert dit laatste argument ons een gewogen conclusie over de validiteit van de toetsscores voor de voorgeschreven interpretatie en het gebruik.

Om deze benadering uit te bereiden van validiteit naar kwaliteit wordt een derde fase toegevoegd. In deze evaluatiefase worden drie criteria geïntroduceerd. Eén criterium dat erop gericht is het interpretatief argument te evalueren, één criterium voor het bewijs dat gepresenteerd is en één criterium voor het validiteitsargument. Deze uitbreiding is in detail beschreven in Hoofdstuk 2 van dit proefschrift. In dit hoofdstuk is de aangevulde benadering geïllustreerd met een bestaand assessment: een competentie assessment voor rijvaardigheid. Dit voorbeeld laat zien dat het mogelijk is om met de aanvullingen op de argumentgerichte benadering van validiteit ook voor evaluatie doeleinden te gebruiken.

De argumentgerichte benadering van validiteit is veelal beschreven in een context van één toets of assessment. In onderwijskundige toepassingen worden toetsen of assessments echter vaak gecombineerd om tot één beslissing te komen. Dergelijke assessmentprogramma's worden bijvoorbeeld gebruikt om diplomabeslissingen te nemen of om de groei van studenten over tijd vast te stellen. Hoofdstuk 3 beschrijft een uitbreiding van de argumentgerichte benadering voor de validering van meerdere toetsen samen. Deze uitbreiding is geïllustreerd aan de hand van een competentie assessment programma (CAP) voor een opleiding voor Social Workers. Het CAP is gevalideerd in samenwerking met een kwaliteitsmanager van het onderwijsprogramma. Deze case studie laat zien dat de benadering een diepgaande evaluatie van het assessment programma kan faciliteren. Daarnaast laat het zien dat de benadering geschikt lijkt voor het valideren van assessment programma's doordat de benadering enerzijds op een algemeen niveau richting geeft aan valideringsonderzoek, maar daarnaast ook in detail valideringsonderzoek kan sturen.

Na de twee uitbreidingen van de argumentgerichte benadering van validiteit is Hoofdstuk 4 er op gericht om de benadering toe te passen in een complexe situatie. Het hoofdstuk is bedoeld om te laten zien dat het begrip over assessments kan toenemen wanneer de argumentgerichte benadering gebruikt

wordt. Daarnaast laat het hoofdstuk zien dat deze benadering onderzoekers en beleidsmakers kan ondersteunen om te beoordelen of ontwerpkeuzes in een assessmentprogramma bijdragen aan de kwaliteit van de beoogde beslissing over leerlingen. Om dit te doen richt het hoofdstuk zich op een nieuw Nederlands assessmentprogramma voor het meten van rekenen. Er wordt geïllustreerd dat in dit assessmentprogramma de belangrijkste claims de vergelijkbaarheid van de losse componenten van het assessmentprogramma betreffen. In het beschreven onderzoek wordt op basis van data geverifieerd of de claims over vergelijkbaarheid houdbaar zijn.
The argumentgerichte benadering van validiteit is zeer bekend in validiteitstheorieën. Het was echter onduidelijk of onderzoekers deze benadering in de praktijk ook gebruikten wanneer zij valideringsonderzoek uitvoeren. In ons onderzoek hebben we daarom gekeken of onderzoekers die valideringsonderzoek uitvoeren, onderscheid maken in het soort validiteitsbewijs dat zij presenteren wanneer zij toetsen met verschillende toetsdoelen valideren. In Hoofdstuk 5 staat een systematische literatuur studie beschreven met daarin centraal de hypothese dat auteurs van artikelen over valideringsstudies verschillende bewijsbronnen aanleveren afhankelijk van het beoogde gebruik van de toetsen. De literatuurstudie bevat 178 artikelen over valideringsstudies over onderwijskundige toetsen. Alle bewijsbronnen die door auteurs beschreven zijn, zijn geclassificeerd binnen het theoretische model passend bij het interpretatieve argument uit de argumentgerichte benadering. Vervolgens is geanalyseerd of er verschillen optraden in de gepresenteerde bewijsbronnen afhankelijk van het beoogde toetsdoel van de toets die gevalideerd werd. De resultaten laten zien dat dit het geval is voor toetsen die geconstrueerd zijn voor selectiedoeleinden. De validiteit van de selectietoetsen is vaker aangetoond met bewijs over het accuraat kunnen voorspellen van toekomstig gedrag. Voor toetsen die geconstrueerd waren voor andere toetsdoelen werd geen ander bewijs gepresenteerd. De resultaten lieten verder zien dat de meerderheid van de artikelen slechts één of twee bewijsbronnen beschrijven en niet, zoals verwacht, een volledig validiteitsargument geven.

Ontwerpprincipes

Vanuit het theoretisch kader werd geconcludeerd dat de argumentgerichte benadering voor validiteit ook geschikt zou kunnen zijn voor het evalueren van kwaliteit van toetsen. We concludeerden ook dat onderzoekers de rationale van deze benadering van validiteit nog niet altijd gebruikten in hun alledaagse

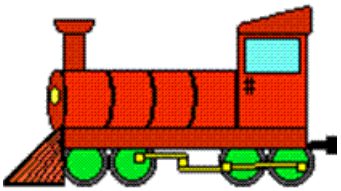
praktijk. Om onderzoekers en toetsontwikkelaars te ondersteunen bij het construeren van de argumenten uit de argumentgerichte benadering van kwaliteit werden in Hoofdstuk 6 ontwerpprincipes opgesteld voor een online beoordelingssysteem. Wanneer onderzoekers door deze software ondersteund worden zijn zij wellicht meer geneigd om verschillende bewijsbronnen te verzamelen en te presenteren wanneer zij de kwaliteit van hun toetsen willen demonstreren.
De ontwerpprincipes die in dit hoofdstuk beschreven staan werden afgeleid uit een vergelijking tussen beschikbare beoordelingsinstrumenten. Het grootste verschil met het nieuwe beoordelingssysteem is dat in dit systeem de argumentgerichte benadering wordt opgenomen. Daarnaast kan het systeem tijdens verschillende fasen van het toetsconstructieproces gebruikt worden. Het derde principe beschrijft dat het beoordelingssysteem erop gericht moet zijn het beoordelingsproces te simplificeren. Dit zou er toe kunnen leiden dat de oordelen van verschillende beoordelaars ook meer vergelijkbaar worden. Ten slotte is geformuleerd dat het beoordelingssysteem andere beschikbare systemen moet includeren om te voorkomen dat het een extra systeem naast de bestaande systemen wordt.

Prototype

Hoofdstuk 7 presenteert een online beoordelingssysteem voor de kwaliteit van toetsen: de QEA (quality evaluation application). De QEA is ontwikkeld in verschillende iteraties van een ontwerpgericht onderzoek waarbij de ontwerpprincipes leidend waren voor de ontwikkeling van een prototype. In dit hoofdstuk is de ontwikkeling van de software beschreven evenals twee evaluatiestudies waarin werd onderzocht of aan de ontwerpprincipes voldaan is. Het hoofdstuk sluit af met de conclusie dat de software vooral veelbelovend is om discussies tijdens het toetsconstructieproces te faciliteren. Daarnaast lijken het prototype en de interface op dit moment voldoende bruikbaar om een nieuwe fase in het onderzoeksproject te starten waarbij de software gebruikt kan worden door potentiele gebruikers.

Dankwoord

Vanaf 2008 heb ik bij Cito mogen werken aan een promotieonderzoek waarvan dit proefschrift het resultaat is. Ik ben hiervoor zowel Cito, als het RCEC, veel dank verschuldigd. Enerzijds omdat ik de tijd, middelen en aanmoedigingen kreeg om 'het treintje' verder te ontwikkelen, anderzijds omdat ik zo vaak in de gelegenheid gesteld ben om het treintje in buitenlanden het scherm in te laten rijden. Het ontstaan van het treintje, de opgelopen vertraging en de uiteindelijke aankomst waren uiteraard niet alleen mijn verdienste, daarom wil ik hier graag nog een aantal mensen persoonlijk bedanken.



Theo,

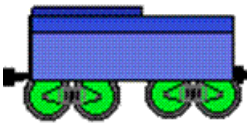
Ik kan me je eerste begeleidingsmoment nog goed herinneren. Je vroeg me of ik het aandurfde om dit met jou, als team, tot een goed einde te brengen. Ik denk dat dat nu gelukt is. En als ons team leek te ontsporen konden we dat het best oplossen door in verre buitenlanden tot diep in de nacht met een biertje een goed gesprek te voeren. Dank voor al je begeleiding, steun, verbale en non verbale feedback, sturing en beschikbaarheid. Maar vooral bedankt dat je me in de laatste jaren een levensles hebt geleerd door te laten zien dat een goede balans tussen werk en privé betekent dat als het aan één van beide kanten tegen zit je support aan de andere kant krijgt en nodig hebt.

Anton,

Toen je me onderweg naar een bepaalde grote kuil (Grand Canyon) als projectleider Referentiesets vroeg wisten we beiden niet dat het zo'n omweg voor de trein zou zijn. Maar dat het me veel heeft gebracht moge duidelijk zijn. Ik besef hoeveel kansen je voor me gecreëerd hebt, dat ik in de combinatie van onderzoeker en projectleider in sneltreinvaart gegroeid ben als wetenschapper en als professional. Zonder jouw onvoorwaardelijke steun, tijd, coaching, reflectie en (wereldse) koffiemomenten was dat nooit gelukt. En gelukkig hoeven we niet te praten, anders had het nog langer geduurd.

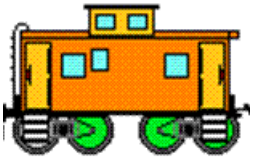
Piet,

Tijdens mijn stage en het begin van mijn promotietraject heb je me oneindig veel geleerd. Dat ik voor mijn kennismakingsgesprek 'psychometrie' moest googlen was snel gerepareerd. De urenlange gesprekken waren de meest waardevolle colleges die ik ooit had. Bedankt dat je me zoveel leerde en me het vertrouwen gaf om als eerste aio bij het RCEC te starten. Dat jij het RCEC oprichtte is niet alleen voor mij belangrijk geweest maar ook voor het vakgebied waar ik inmiddels onderdeel van uit maak.



Robert en Matthieu,

Jullie zijn het ideale paranimf-duo, dat blijkt wel uit het feit dat jullie de rol jarenlang, zonder klagen, met veel toewijding hebben vervuld. Bedankt voor de eindeloze reflectie, intervisie en sloten koffie. Bedankt voor de vriendschap, support en alle discussies over relevante en minder relevante levensvragen. Maar ook bedankt dat ik de kunst van het promoveren bij jullie mocht afkijken zodat het bij mij net iets makkelijker gaat. Ik ben er trots op dat we er uiteindelijk allemaal in slaagden ons onderzoek tot een goed eind te brengen.



Pokkers,

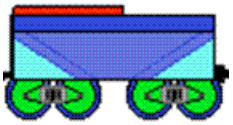
Je zou kunnen zeggen dat ik bij POK opgegroeid ben. Ik kwam er binnen als stagiair en mocht er ruim 6 jaar als jonge onderzoeker leren en werken. De cultuur binnen de afdeling heb ik altijd gewaardeerd. De inhoudelijke discussies, werklunches met kroketten, collega's die de dag uitroken om aan hun gezondheid te werken en spraakmakende kerstdiners. Maar vooral jullie interesse in mijn onderzoek, in mijn projecten en in mij als persoon hebben er voor gezorgd dat ik me altijd thuis voelde bij POK.

Patricia en Servaas,

Bedankt dat jullie het blinde vertrouwen in 'mijn team' altijd hebben waargemaakt. Zonder jullie inzet tijdens de refsets, de eeuwige spirit en het no nonsense problemen oplossen had ik nooit daarnaast ook nog onderzoek kunnen doen. Dank voor het luchtig houden van de grootste crises, voor het verzenden van post aan kraakpanden en de SvZjes die nooit doorgingen.

Anke, Hendrik en Marie-Anne,.

Ik ben blij dat jullie de oranje bank vaak willen delen voor advies, afstemming of om het weekend te bespreken. Bedankt dat jullie me af en toe afleiden, bijpraten over de laatste stand van zaken of gewoon even heel hard 'dat kan toch niet' roepen.



Cito-collega's,

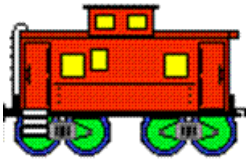
Tijdens mijn onderzoek hebben ontelbaar veel Cito-collega's belangstelling getoond voor mijn proefschrift, daar kan ik alleen maar heel dankbaar voor zijn. Meer specifiek, ben ik veel dank verschuldigd aan Mark Volmer en Patrick de Klein; bedankt dat jullie mijn kleurrijke kwaliteitstool een make-over hebben gegeven. Collega's van CI, bedankt voor jullie reisgezelschap en collegialiteit. Rianne en Bernadette, bedankt voor de fijne samenwerking in onze nieuwe rollen, ik kijk er naar uit veel mooie dingen met elkaar te maken.

RCEC collega's,

Ook de collega's en mede aio's van het RCEC hebben een belangrijke bijdrage geleverd aan dit proefschrift. Maarten, Maaike, Jorine, Sebastiaan, Hiske, Britt en Fabienne, dank voor jullie input tijdens dit traject. Bernard, Birgit en Lorette, bedankt dat jullie altijd je best hebben gedaan om mij ook onderdeel van de UT te laten zijn en daar alles voor te regelen als het nodig was. En Dorien, bedankt dat je samen met mij in no-time een hele literatuurreview opnieuw deed. Het was leuk, verfrissend en motiverend om daar samen aan te werken.

Internationale collega's,

Gedurende dit onderzoek heb ik heel erg vaak op congressen en in internationale verbanden mijn onderzoek mogen presenteren. De suggesties die ik tijdens deze gelegenheden van collega's kreeg waren van grote waarde en zonder deze inhoudelijke discussies had ik dit proefschrift niet kunnen schrijven. Ik wil Liesbeth Baartman, Chad Buckendahl, Michael Kane, Paul Newton, Stuart Shaw, Gordon Stobart en Allistair Pollitt hartelijk danken voor hun interesse en feedback.

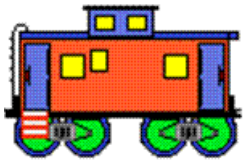


(Trein)vrienden,

Ik heb het geluk om te mogen werken met mijn vrienden. Dat zorgt ervoor dat er regelmatig tijd is voor mibo's, er meestal goed gezelschap in de trein is en er altijd genoeg onderwerpen voor discussie zijn op het terras. Arjan, Marieke, Roos, Vera, Tjeerd Hans, dankjewel dat jullie naast collega's ook vrienden zijn. Marcel, bedankt dat je voor mij altijd consignatiedienst draait. Henk, als origineel lid van de treinvrienden ben ik blij dat je voor mij een trein hebt kunnen tekenen. Bedankt voor al je werk aan de omslag en de uitnodiging. Gelukkig zijn er ook mensen in mijn leven die niet over toetsen praten: Georgia, Karan, Dylan, wat fijn dat jullie me af en toe herinneren aan de wereld buiten Cito.

Familie,

Syl, dankjewel dat je me op je eigen manier en vol humor jong houdt en inspireert. Iedereen zou jaloers moeten zijn op zo'n zusje. Papa, bedankt dat je me geleerd hebt zelfstandig te zijn, het komt me dagelijks goed van pas. Mama en Cees, bedankt voor jullie support en steun tijdens mijn studiejaren. Ook mijn andere familie, de Molenaars, Hans, Marijke, Cobi, Afke, dank voor jullie goede raad tijdens de familie-koffie en dat jullie me altijd het gevoel geven dat ik er bij hoor. Dat geldt voor de hele familie maar natuurlijk het meest voor Sytze, Marloes, Nienke, Menno, Margot en Yinthe.



Mark,

Bedankt dat je mijn alles bent en dat bedoel ik letterlijk: vriendje, maatje, collega, Uteregs liefie, helpdesk, reviewer, co-presenter, datapunt, reisgenoot, koffiezetter, probleemoplosser, muizenvanger, huisman, opruimer en nog veel meer. Je bijdrage aan dit proefschrift is onbeschrijfelijk, zonder jou was het er niet. Bedankt dat je altijd in me geloofde, dat je zo trots bent, en dat je er soms gewoon een tijdje niet naar vroeg. Wat heerlijk dat het nu af is.