Chapter 9

Towards a Comprehensive Evaluation System for the Quality of Tests and Assessments

Saskia Wools

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.

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

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 1 Comparison of standards, guidelines, and evaluation systems

	ITC	AERA Standards	AEA- Europe	ETS	Cambridge Assessment	JCTP	COTAN	EFPA	Baartman
Purpose									
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	
Intended audience									
Test specialists		x	x				x	x	
Teachers	x		x			x			x
Users		x	x			x	x		x
Companies			x	x	x				
Object of evaluation									
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		

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

Table 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 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>
- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research 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.
- Educational Testing Service (ETS). (2002). *ETS standards for quality and fairness*. Princeton, NJ
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: NIP.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices.
- 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 <http://www.efpa.eu/professional-development/tests-and-testing>

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