cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Developing questionnaires to measure students' expectations and perceptions of assessment quality

Karin J. Gerritsen-van Leeuwenkamp[1,2]*, Desirée Joosten-ten Brinke[1,3] and Liesbeth Kester[4]

Received: 11 February 2018
Accepted: 10 April 2018
First Published: 16 April 2018

*Corresponding author: Karin J. Gerritsen-van Leeuwenkamp, Saxion University of Applied Sciences, PO Box 70.000, 7500 KB Enschede, The Netherlands
E-mail: karin.gerritsen-vanleeuwenkamp@ou.nl

Reviewing Editor:
Sammy King Fai Hui, The Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

Abstract: Students form expectations and perceptions about assessment quality, and they experience the consequences of inferior assessment quality. These expectations and perceptions influence how students use assessment and the information it provides in educational practice. Unfortunately, little is known about students' expectations and perceptions of assessment quality, and no evaluation tool is available to measure students' expectations and perceptions of assessment quality to obtain more insight into their perspectives. Therefore, this study aims to construct two questionnaires: the Students' Expectations of Assessment Quality Questionnaire (SEAQQ) and the Students' Perceptions of Assessment Quality Questionnaire (SPAQQ). Both questionnaires were analysed using principal axis factoring with direct oblimin rotation of the data (213 higher education students). This resulted in 39 items spread over six factors: (1) effects of assessment on learning, (2) fairness of assessment, (3) conditions of assessment, (4) interpretation of test scores, (5) authenticity of assessment, and (6) credibility of assessment. Cronbach's alphas varied from α = .76 to α = .94. Educational organisations can use these questionnaires to evaluate assessment quality from a students' perspective to improve assessment practice.

ABOUT THE AUTHORS

Karin J. Gerritsen-van Leeuwenkamp is a PhD-student at the Welten Institute of the Open University of the Netherlands and educational advisor at the Quality Assurance Office of Saxion, University of Applied Sciences. Her expertise includes quality of testing and assessment, students' perspective on assessment quality, and the relation between assessment and learning.

Desirée Joosten-ten Brinke is an associate professor in testing and assessment at the Welten Institute of the Open University of the Netherlands and the Teacher training Institute at Fontys, University of Applied Sciences. Her expertise includes formative and summative assessment, assessment of prior learning, e-assessment, and quality of testing and assessment.

Liesbeth Kester is full professor Educational Sciences at the Department of Education & Pedagogy at Utrecht University. Her expertise includes multimedia learning, hypermedia learning, personalised learning, cognitive aspects of learning, and designing and developing flexible learning environments.

PUBLIC INTEREST STATEMENT

Higher education students undertake many assessments. When the quality of assessments is inferior, the information they provide might be inaccurate. Therefore, educational organisations invest time and effort into assuring assessment quality. However, if students do not perceive assessment as transparent or reliable, they will not use assessment and its information properly in educational practice. To attain assessment quality, educational organisations should measure the students' perspectives. Two questionnaires were constructed, the Students' Expectations of Assessment Quality Questionnaire and the Students' Perceptions of Assessment Quality Questionnaire. The validation process confirmed that six scales capture students' expectations and perceptions of assessment quality: (1) effects of assessment on learning, (2) fairness of assessment, (3) conditions of assessment, (4) interpretation of test scores, (5) authenticity of assessment, and (6) credibility of assessment. Each scale showed sufficient reliability. Educational organisations can use these questionnaires to evaluate assessment quality from a students' perspective to improve assessment practice.

Subjects: Education - Social Sciences; Testing, Measurement and Assessment; Test Development, Validity & Scaling Methods; Perception; Educational Research; Higher Education; Assessment; Assessment & Testing

Keywords: assessment quality; perception; expectation; scale development/testing; questionnaire; higher education; evaluation; validation

1. Introduction

Higher education students undertake many tests and assessments throughout the course of their studies. These may include knowledge tests, classroom questioning, performance, and portfolio assessments. Information from assessment will be used for the purposes of learning, selection and certification, and accountability (Anderson & Rogan, 2010; Stobart, 2008). However, when the quality of assessment is inferior, the information obtained may be inappropriate and/or inaccurate. Therefore, educational organisations invest time and effort in assuring, monitoring, and improving the quality of several aspects of assessment (e.g. test items, tasks, assessment programmes, assessment organisation, and assessment policy).

In higher education, attaining assessment quality requires more than test developers assuring the technical quality of an individual test based on psychometric criteria (Linn, Bakker, & Dunbar, 1991). Although standardised tests are used in assessment programmes, these are combined with more authentic, interactive, integrated, and flexible forms of assessments (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Birenbaum et al., 2006; Shepard, 2000). For example, performance tasks are used in which authentic problems are assessed in several real-life contexts (Shepard, 2000). The role of individuals, and their interactions with each other and the context, becomes more important (Moss, Pullin, Gee, & Haertel, 2005). Furthermore, assessment is an essential part of learning and teaching (Moss et al., 2005) because it collects information that students and teachers can use to evaluate student achievement and to alter and improve the teaching and learning process (McMillan, 2007). Therefore, the quality of assessment should be evaluated on its fitness for purpose, its proposed use, and its consequences (Baartman et al., 2006; Boud, 2000; Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, & Kester, 2017; Van der Vleuten et al., 2012).

Although, assessment developers can provide important information about the quality of assessment, their perspectives of that quality should be combined with the perspectives of other stakeholders (i.e. students, teachers, and employers). Firstly, the perspectives of stakeholders are important because they use assessment in educational practice, which influences its quality (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007a). For example, when students are not fully engaged in assessments because of disinterest this might result in a low assessment score, which might better represent their motivation level than their mastery level of the learning content. This affects the appropriateness of the interpretations that are based on those scores (Kane, 2013; Wise & Cotten, 2009; Zilberberg, Brown, Harmes, & Anderson, 2009). Furthermore, some teachers might give higher grades to students who have performed well in the past based on their prior experiences with that student instead of only assessing a student's current performance (halo bias). This influences the reliability of the assessment score (Archer & McCarthy, 1988; Dennis, 2007). Secondly, stakeholders have perspectives on the purpose and use of assessment, and they experience the consequences of inferior quality (Baartman et al., 2007a; Gulikers, Biemans, & Mulder, 2009; Moss, 1994). As Harvey and Green (1993) explained: "we might all have different understandings of quality in higher education and that none of us is necessarily wrong or right does not mean, however, that we are absolved of the responsibility for maintaining and enhancing quality" (p. 29). Therefore, while attaining assessment quality, the stakeholders' perspectives of assessment quality should be considered, and differences and contradictions should be discussed (Harvey & Green, 1993). Only then is it possible to obtain better insight about the stakeholders' compliance with assessment, their ideas of assessment, their use of assessment, and, thus, whether assessment is fit for educational practice (Dijkstra, Van der Vleuten, & Schuwirth, 2010).

Students are one group of stakeholders that should be considered. However, their perspectives are not always acknowledged (Levin, 2000). This is remarkable because students are stakeholders in an educational organisation and, according to the democratic structure of higher education, they have their own rights and responsibilities, and their voices should be heard (Levin, 1998; Svensson & Wood, 2007). For example, in Europe, the Bologna Process adopted a governance approach in which students are required to participate in the process of quality assurance of higher education at the European, national, and institutional levels (EHEA, 2014; ENQA, 2009; Klemenčič, 2012). Moreover, students have unique information and perspectives (Levin, 2000) because they directly experience the consequences of inferior assessment quality. The perspectives of students differ from that of teachers, managers, and employers (Gulikers et al., 2009; Meyer et al., 2010; Van de Watering & Van de Rijt, 2006). As Brown, McInerney, and Liem (2009) stated: “it seems self-evident that students perceive, respond to, and shape what assessment means and that their internal evaluations of and attitudes towards assessment are not necessarily in line with those of their parents, teachers, or societies” (p. 4). For example, a study by Meyer et al. (2010) found that teachers disagree with students about the unfairness of assessment results, and Van de Watering and Van de Rijt (2006) found that while students overestimated the item difficulty, teachers underestimated it.

Students’ perspectives on assessment quality must be understood in light of their expectations and perceptions. These two constructs are important because students’ expectations and perceptions guide their behaviour in the learning environment (Roese & Sherman, 2007; Struyven, Dochy, & Janssens, 2005). Perceptions refer to how people understand and assign meaning to the information in the environment (Zimbardo, Weber, & Johnson, 2009). Expectations influence this process by determining, and sometimes even biasing, what information people see in the environment and how this information is interpreted (Roese & Sherman, 2007; Zimbardo et al., 2009). Expected information is easier and more efficient to process and understand (Roese & Sherman, 2007). In contrast, unexpected information is surprising; therefore, it is more carefully processed (Roese & Sherman, 2007). Changes in behaviour occur when there is a discrepancy between the perceived information and someone’s expectations. For example, Könings, Brand-Gruwel, van Merriënboer, and Broers (2008) found that students’ decreased expectation to perception scores of the study environment were related to less motivation, decreased use of deep processing strategies, and increased anxiety about failing. In contrast, when the perceived information and expectations are fully aligned, no changes in behaviour occur (Roese & Sherman, 2007).

Unfortunately, little is known about students’ expectations and perceptions of assessment quality. According to Ecclestone (2012), student expectations focus mainly on the quality of the selection and certification function of assessment. Some studies have investigated students’ perceptions of assessment quality criteria, such as authenticity, fairness, and item difficulty (Gerritsen-van Leeuwenkamp et al., 2017). When students do not perceive an assessment as being authentic, their learning could be obstructed (Gulikers, Kester, Kirschner, & Bastiaens, 2008). Moreover, students perceive assessments as fairer and more educationally valuable when assessments are embedded in realistic contexts, when hard work pays off, when the assessment provides useful feedback, and when grading is consistent (O’Donovan, 2016; Sambell, McDowell, & Brown, 1997). However, no evaluation tool is available to measure students’ expectations and perceptions of assessment quality to obtain more insight into their perspectives. Therefore, the present study aimed to construct two questionnaires: the Students’ Expectations of Assessment Quality Questionnaire (SEAQQ) and the Students’ Perceptions of Assessment Quality Questionnaire (SPAQQ).

2. Method

2.1. Phase 1: questionnaire development

A literature review operationalised assessment quality and based on objective text analyses 98 assessment quality criteria were clustered in four subdimensions (Gerritsen-van Leeuwenkamp et al., 2017). Firstly, validity refers to the degree to which the evidence and the arguments support the interpretation and inferences based on test scores (Kane, 2001). Secondly, transparency refers to the degree to which stakeholders understand assessment (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007b). Thirdly, reliability refers to the degree of consistency, reproducibility, and generalisability of assessment scores (Downing, 2004; Schuwirth & Van der Vleuten, 2004). Fourthly, 'other' consists of assessment quality criteria that do not belong to any of the other three subdimensions. A questionnaire item was formulated for each of the 98 assessment quality criteria, based on the retrieved text fragments of the 78 journal articles that were included in the literature review (Gerritsen-van Leeuwenkamp et al., 2017). The guidelines for question construction were taken into account, e.g. jargon, lack of clarity, and ambiguity were avoided (Creswell, 2008). To limit the length of the questionnaire, criteria were clustered and duplicates were removed whenever possible. This resulted in 40 items. Table 1 provides insights into the relationships between the assessment quality criteria of the literature review and the items in the SEAQQ and SPAQQ. For validation purposes, one open-ended question asked students to describe their expectations of assessment quality and one open-ended question asked for their perceptions of assessment quality. One closed item asked students to assess their own overall expectations and one closed item asked students for their overall judgment of their perceptions of assessment quality.

2.2. Phase 2: pilot

In the pilot phase, 29 students answered the questionnaires. They evaluated the comprehensibility of the items ("did you have doubts about the meaning of some items?"), the completeness of the questions ("if you are thinking about assessment quality, do you miss important items?"), and the attractiveness of the questionnaire ("is the questionnaire inviting to complete?"). For this pilot, a purposive sample was taken of a total of 57 first- and second-year students in the Bachelor of Nursing and the Bachelor of Podiatry programmes at a university of applied sciences in the Netherlands. Two weeks before they were scheduled to answer the questionnaires, the students were informed by e-mail about the purpose of the study and the research process. Participation in this pilot phase was included in the students' schedules. At the scheduled meeting, the students were again informed of the purpose of the study and the research process, and then the students (n = 16) completed the digital questionnaires in the presence of the researcher. Students (n = 41) who did not attend the meeting were asked by e-mail to complete the questionnaires at home and return it within one week. A total of 30 students responded to the questionnaires. One respondent was excluded from further analysis because, by his own admission, he did not answer the questionnaires seriously. Thus, 29 students (25 females, 4 males, M_{age} = 20.52, age range 18–30 years) participated. Eighteen of the 40 items, which operationalised assessment quality, received remarks on their comprehensiveness (13 items by one student; 3 items by two students; 1 item by three students; and 1 item by four students). After considering the open feedback and the scores of the items, 14 items were reformulated, and some terms were further clarified. All the students understood the difference between expectations and perceptions. According to 79.3% of the students, no essential items were missing. Three items were added based on the students' feedback. A total of 79.3% of the students thought that the questionnaires were inviting to complete, and 75.9% of the students noted that the number of items was acceptable. The font size was approved by 93.1% of the students. Furthermore, the character limit for the open question text boxes was removed at the request of some students, and the text boxes were enlarged. All adaptations were made in both the SEAQQ and the SPAQQ.

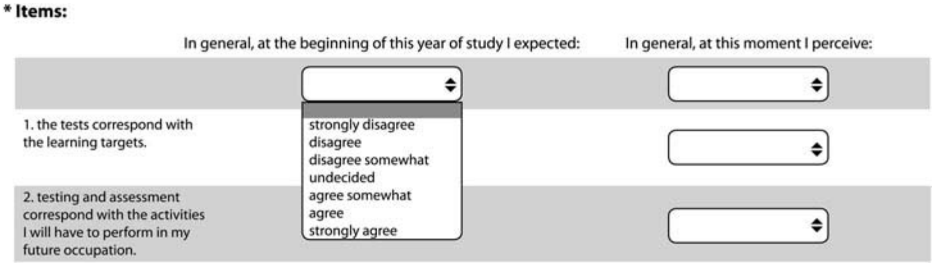
The final versions of the SEAQQ and SPAQQ consisted of 43 closed items that operationalise assessment quality, one closed item to assess the students' overall expectations/perceptions and one open question for validation purposes on each questionnaire. The closed items allowed responses

Table 1. Relationships between the assessment quality criteria of the literature review (Gerritsen-van Leeuwenkamp et al., 2017) and the items in the SEAQQ and the SPAQQ

Scale validity (V)		Scale transparency (T)		Scale reliability (R)		Scale other (O)	
Item	**AQC**	**Item**	**AQC**	**Item**	**AQC**	**Item**	**AQC**
V01	acceptability	T01	comparability	R01	consistency	O01	difficulty
V02	content-validity		equity		objectivity	O02	challenge
	fairness		fairness		reproducibility	O03	supportive
	representativeness	T02	fairness	R02	consistency	O04	supportive
	validity	T03	confidence	R03	accuracy		
V03	consequences		defensible		precisely		
	encouraging		effectiveness		reliability		
	validity		trustworthiness		reproducibility		
V04	consequences	T04	engaging	R04	manageability		
	validity	T05	motivation		timely		
V05	adequacy	T06	authenticity	R05	manageability		
	construct-validity		fidelity	R06	soundness		
	validity		interactiveness	R07	robustness		
V06	concurrent	T07	accessibility		security		
	criterion-validity		transparency	R08	accuracy		
	predictive		well documented	R09	soundness[1]		
	validity	T08	self-assessment				
V07	affordability	T09	sustainable				
	worthy	T10	challenge				
	meaningfulness		supportive				
V08	generalisability		sustainable				
	relevance	T11	sustainable				
V09	generalisability	T12	self-assessment				
V10	meaningfulness	T13	authenticity				
	recognisable	T14	complexity				
V11	systematically	T15	self-assessment				
		T16	transferability				
		T17	efficiency				
			feasibility				
			reasonable				
		T18	feasibility[1]				
		T19	equity[1]				

Note: Item is the item code of the SPAQQ and SEAQQ. AQC is the assessment quality criterion. [1] Items added based on feedback of students.

Figure 1. Example of the matrix structure of the questionnaires.



*** Items:**

In general, at the beginning of this year of study I expected:

In general, at this moment I perceive:

1. the tests correspond with the learning targets.

2. testing and assessment correspond with the activities I will have to perform in my future occupation.

strongly disagree
disagree
disagree somewhat
undecided
agree somewhat
agree
strongly agree

on a 7-point scale (ranging from 1 = completely disagree to 7 = completely agree). The closed items of the SEAQQ and SPAQQ were presented to the students in a matrix. The items were listed in the rows of the matrix, and the drop-down menu of answers was listed in the columns (see Figure 1).

2.3. Phase 3: validation

In order to evaluate whether the scores of the questionnaires reflect the two constructs, a validation process took place.

2.3.1. Participants

A total of 769 students in the Bachelor of Podiatry, Bachelor of Nursing, Bachelor in Art and Technology, Bachelor of Teacher Training for Secondary Education in English, and Bachelor of Teacher Training for Secondary Education in Spanish from two universities of applied sciences in the Netherlands were invited to participate in this study. The response was 222 (28.9%), of which nine students were excluded (five minor students were excluded because no parental consent form was returned; two students were excluded because they noted that they occasionally interpreted the answer scale incorrectly; two students were excluded because they noted that they did not respond to the questionnaire seriously). Thus, 213 students (191 females, 22 males, M_{age} = 23.61, age range 17–60 years) participated in this study. The participants were entered into a raffle, and two gift vouchers of 20€ were awarded.

2.3.2. Data collection procedure

The participants were informed about the research purpose and process by e-mail at least two weeks before their scheduled participation in the study. When feasible, the questionnaires were administered to groups of students in the presence of the researcher; students that were absent that day received a reminder to answer the questionnaires at home. The students received an e-mail invitation to respond to the questionnaires, and after the first invitation, they received three reminders at weekly intervals. A consent form was included with the questionnaire. Permission to participate was requested from the guardian or parent of minor students by e-mail and by letter. For each item, students had to first recall their expectations of assessment quality at the start of the academic year, and then they had to evaluate their current perceptions of assessment quality after approximately 30 weeks.

2.3.3. Data analysis procedure

Confirmative factor analysis, in which the factor structure is determined beforehand, was considered but not used in this study. Although four subscales were identified based on a literature review, this review did not focus solely on students' perspectives. Therefore, it was decided that it was better to explore the data. Exploring the underlying factor structure in the data on students' expectations and perceptions of assessment quality rather than simply confirming it reveals the differences between the factor structure and the subscales based on previous literature. Principal axis factoring (PAF) was performed (Field, 2013). Oblique rotation, in the form of direct oblimin rotation, was chosen as the factor rotation method because correlations between factors were expected (Field, 2013). After factor analysis, a reliability analysis was performed. When necessary, this process was repeated until reliable factors were achieved (Field, 2013). After determining the factor structure of the SEAQQ and SPAQQ separately, the two structures were compared, and one common factor structure was chosen for both questionnaires to enhance the comparability of both constructs. Furthermore, reliability analyses were performed on this structure, and reliability analyses were performed on three subscales from the literature (validity, transparency, and reliability) to identify the differences.

The first author coded the responses to the two open questions using the codes of the items of the questionnaires. Maxqda version 11 data-analysis software was used for the coding process. The second author evaluated the coded fragments. This was done to ensure accurate coding of the fragments; the second author either approved the coding choices or challenged them by offering an alternative code (Carroll, Booth, & Cooper, 2011). According to the second author, 95.8% (SEAQQ) and 97.5% (SPAQQ) of the text fragments were coded accurately.

3. Results

3.1. Explorative factor analysis

The data (N = 213) of the SEAQQ and SPAQQ were analysed separately using PAF with direct oblimin rotation. Before executing the PAF for each questionnaire, it was determined whether it is likely that there are any factors underlying the measured variables (factorability) (Brace, Kemp, & Snelgar, 2016). Most of the correlation coefficients were > .3 in both correlation matrices, and the majority of these correlations appeared to be significant. This indicates the presence of a factor structure under the variables (Brace et al., 2016). Bartlett's test for sphericity indicated that the SEAQQ and SPAQQ data were probably factorable, $p < .05$. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, which indicates the amount of variance within the data that can be explained by factors (Brace et al., 2016), was good (SEAQQ KMO of .95; SPAQQ KMO of .91). Furthermore, the Measure of Sampling Adequacy (MSA) for each variable indicated that none of the items needed to be dropped (Brace et al., 2016) (the SEAQQ MSA values ranged between .84 and .97; the SPAQQ MSA values ranged between .82 and .96).

The PAF of the SEAQQ identified seven factors, these represent: effects of assessment on learning (8 items, α = .93), fairness of assessment (4 items, α = .80), conditions of assessment (9 items, α = .93), credibility of assessment (9 items, α = .90), authenticity of assessment (4 items, α = .78), and feedback (2 items, α = .88). One factor could not be meaningfully interpreted; it contained one item (T17) with a loading only on this factor. Six items had no loading > .3. After removing items T12 and T19, the reliability of the related factors improved (credibility of assessment: 8 items, α = .91; authenticity of assessment 3 items, α = .82). The factor analysis was repeated to make sure the factor structure still held (Field, 2013). This PAF resulted in a new factor structure. The results showed six factors, these represent: effects of assessment on learning (11 items, α = .94), fairness of assessment (5 items, α = .81), credibility of assessment (6 items, α = .88), conditions of assessment (10 items, α = .94), and authenticity of assessment (5 items, α = .82). One factor could not be meaningfully interpreted; it consisted of two items with a loading only on this factor (R03; T15). Two items had no loading > .3. The reliability coefficients indicated a sufficient degree of reliability for each factor (Brace et al., 2016).

The PAF of the SPAQQ identified nine factors. Two items (T12 and T19) were removed in accordance with the outcome of the PAF of the SEAQQ because the goal was to establish one factor structure for both the SEAQQ and SPAQQ. Five of the factors in the SPAQQ were comparable to the SEAQQ. The factors represent: effects of assessment on learning (7 items, α = .88), fairness of assessment (3 items, α = .78), authenticity of assessment (6 items, α = .80), feedback (2 items, α = .84), conditions of assessment (6 items, α = .84), interpretation of test scores (5 items, α = .77), and credibility of assessment (4 items, α = .68). Two factors could not be meaningfully interpreted; they contain one or two items with a loading only on this factor. Five items had no loading > .3. The reliability coefficients indicated a sufficient degree of reliability for each factor (Brace et al., 2016).

After interpreting the factor structures of the PAFs for the SEAQQ and the SPAQQ separately (see Table 2), the results were compared in order to establish one factor structure. The factor structure

Table 2. Factor loadings of the pattern matrix of the PAF of the SEAQQ and the PAF of the SPAQQ

Item	Factor effects of assessment on learning		Factor fairness of assessment		Factor conditions of assessment		Factor interpretation of test scores		Factor authenticity of assessment		Factor credibility of assessment		Factor feedback	
	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ
V01*							-		−.36	.36	**.44**		-	
V02*			**.45**	**.37**			-						-	
V03*	**.64**	**.69**					-						-	
V04							-				**.49**		-	
V05					−.36		-	**.38**			.33		-	
V06							-	**.55**			.38		-	
V07*	**.62**	**.32**					-						-	
V08					−.32		-	**.73**					-	
V09	-	-	-	-	-	-	-	-	-	-	-	-	-	-
V10*	**.63**	**.49**					-						-	
V11*					**−.64**	**−.46**	-						-	
T01*			**.81**	**.75**			-						-	
T02*			**.65**	**.65**			-						-	
T03*	.36	.43				−.45	-				**.35**		-	
T04							-				**.39**		-	
T05*	**.74**	**.76**					-						-	
T06*							-		**−.36**	**.67**			-	
T07*							-		**−.33**	**.30**			-	
T08	**.43**						-		−.34				-	
T09	**.48**					−.39	-		−.31				-	
T10	**.63**						-			.45			-	
T11*	**.68**	**.70**					-						-	
T12	-	-	-	-	-	-	-	-	-	-	-	-	-	-

(Continued)

Table 2. (*Continued*)

Item	Factor effects of assessment on learning		Factor fairness of assessment		Factor conditions of assessment		Factor interpretation of test scores		Factor authenticity of assessment		Factor credibility of assessment		Factor feedback	
	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ	SEAQQ	SPAQQ
T13*							-		**−.65**	**.61**			-	
T14							-	.34	**−.56**				-	
T15	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T16*							-		**−.55**	**.48**			-	
T17			**.30**				-						-	
T18					**−.38**		-					.63	-	
T19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R01*				.36			-				**.72**	**.35**	-	
R02*				.32			-				**.62**	**.32**	-	
R03							-	**.44**					-	
R04*	.45				**−.30**	**−.32**	-						-	
R05*					**−.56**	**−.43**	-						-	
R06*					**−.44**	**−.74**	-						-	
R07					**−.62**		-					.33	-	
R08					**−.31**		-						-	
R09*					**−.63**	**−.79**	-						-	
O01			**.51**				-						-	
O02		**.37**					-						-	
O03	**.38**		.39				-						-	.93
O04	**.43**						-				.36		-	.81

Note: Factor loadings < .3 are not listed. The codes of the items V (validity), T (transparency), R (reliability), or O (other) refer to the four subscales based on a literature review.
Of the 43 items shown in this table:

- Twenty-one items are marked with an asterisk because they have comparable classifications in the PAF of the SEAQQ and of the SPAQQ.
- Eighteen items are classified differently in the PAF of the SEAQQ and the PAF of the SPAQQ.
- The preferred classification, based on a comparison of the PAF of the SEAQQ and the PAF of the SPAQQ, is in bold.
- Four items were deleted in both questionnaires (V09, T12, T15, and T19).

was established by looking at the similarities and differences between the PAFs of the SEAQQ and SPAQQ. As shown in Table 2 and marked with an asterisk, 21 items loaded on the same factors in both PAFs. Of those items, two did not substantially match the factor; therefore, they were manually assigned to another factor (in accordance with the PAF of the SEAQQ). Eighteen items loaded differently in the two PAFs. These items were manually linked to a factor (13 items according to the PAF of the SEAQQ and 5 according to the PAF of the SPAQQ), based on the following three arguments. First, the item had to substantially match the factor. For example, the factor 'feedback' in the SPAQQ was matched to the factor 'effects of assessment on learning', because feedback has an effect on students' learning. Second, items with a meaningless loading in one of the PAFs were assigned to the other PAF. For example, item R03 had a loading in the PAF of the SEAQQ on a factor that could not be meaningfully interpreted. In the PAF of the SPAQQ, this item had a meaningful loading under the factor 'interpretation of test scores', so it was linked to the factor according with the PAF of the SPAQQ. Third, items with an insufficient loading in one of the PAFs were assigned to the other. For example, item O02 had no loading > .3 in the SEAQQ; therefore, it was linked to the factor based on the PAF of the SPAQQ. Two items were deleted in both questionnaires to increase the reliability (T12 and T19), and two items were removed because of their loading (T15 and V09). The deletion of these four items had no impact on the validity of the questionnaires because the remaining items still covered the assessment quality criteria defined in the literature (Table 1). As seen in Table 2, the chosen classification is marked in bold. Six factors are distinguished: effects of assessment on learning (11 items, SEAQQ α = .94, SPAQQ α = .89), fairness of assessment (5 items, SEAQQ α = .81, SPAQQ α = .78), conditions of assessment (8 items, SEAQQ α = .92, SPAQQ α = .80), interpretation of test scores (4 items, SEAQQ α = .88, SPAQQ α = .76), authenticity of assessment (5 items, SEAQQ α = .82, SPAQQ α = .76), and credibility of assessment (6 items, SEAQQ α = .89, SPAQQ α = .81). These reliability coefficients indicate a sufficient degree of reliability for each factor (Brace et al., 2016). Table 3 shows the final structure of the 39 items within six factors for the SEAQQ and SPAQQ.

In the SEAQQ, there appears to be a moderate positive significant correlation (r = .52, p < .01) (Brace et al., 2016) between the closed item, "In general, I had a positive expectation of assessment quality at the beginning of the school year', and the mean score of the 39 items on the SEAQQ. The SPAQQ showed a strong positive significant correlation (r = .81, p < .01) (Brace et al., 2016) between the closed item, 'In general at this moment, I have a positive perception of assessment quality', and the mean score of the 39 items on the SPAQQ. These moderate to strong positive correlations indicate that the students" scores on the SEAQQ and SPAQQ can be used for the operationalisation of the constructs: "students' expectations of assessment quality" and "students' perceptions of assessment quality".

A comparison of student comments on the two open questions with the items on the questionnaires showed that the questionnaires cover all the comments given by the students. This indicates that no extra items were required. Table 4 presents examples of the students' comments on the open questions related to the items in the questionnaires.

3.2. Scales based on the literature

Reliability analyses were performed on the three subscales based on previous literature: validity (11 items V01-V11, SEAQQ α = .91, SPAQQ α = .86), transparency (19 items T01-T19, SEAQQ α = .92, SPAQQ α = .87), and reliability (9 items R01-R09, SEAQQ α = .91, SPAQQ α = .82). Two items (V09 and T15) were removed to increase the reliability of the scales for both the SEAQQ and the SPAQQ. The deletion of the two items had no impact on the validity of the scales because the remaining items still covered the assessment quality criteria (Table 1). The removal of the items resulted in valid and more reliable questionnaires (validity 10 items, SEAQQ α = .92, SPAQQ α = .87, transparency 18 items, SEAQQ α = .92, SPAQQ α = .87). These reliability coefficients indicate a sufficient degree of reliability for each factor (Brace et al., 2016).

Table 3. Final structure of the SEAQQ and the SPAQQ with reliability coefficients per factor

Factor	Code item	Number item	Item SEAQQ: In general, at the beginning of this year of study I expected: SPAQQ: In general, at this moment I perceive:	SEAQQ α	SPAQQ α
1. Effects of assessment on learning 11 items	V03	34	Testing and assessment have a positive effect on my learning.	.94	.89
	V07	33	Testing and assessment add value to the time I have spent on the work done.		
	V10	37	Testing and assessment are valuable instances of learning in their own right.		
	T05	38	Testing and assessment motivate me to continue learning.		
	T08	39	Testing and assessment help me to navigate my own learning process.		
	T09	40	Testing and assessment are geared towards the retention of my competencies in the longer run.		
	T10	41	Testing and assessment prepare me well for future learning activities.		
	T11	36	Testing and assessment give me the confidence to continue learning.		
	O02	13	The tests are challenging.		
	O03	30	When I get feedback on tests it shows clearly what I have not yet mastered.		
	O04	31	When I get feedback on tests it shows clearly what I have already mastered.		
2. Fairness of assessment 5 items	V02	1	The tests correspond with the learning targets.	.81	.78
	T01	5	Testing and assessment are the same for all students in my year.		
	T02	7	Testing and assessment are fair.		
	T17	12	Testing and assessment can be done in the time given.		
	O01	4	The difficulty of testing and assessment concur with the level of my education.		
3. Conditions of assessment 8 items	V11	19	The tests and assessments are organised well.	.92	.80
	T18	21	Tests have been spread out evenly during the periods set for testing in the year of study.		
	R04	32	When I get feedback on my tests, I will receive it in time.		
	R05	20	The team of teachers in my educational programme are accomplished in testing and assessment.		
	R06	15	All tests feature correct language.		
	R07	18	During testing and assessments there are no disturbing external factors, such as fraudulent behaviour.		
	R08	24	Whether I pass or fail is based correctly on the score of a test I have taken.		
	R09	16	Tests have been constructed with care.		

(*Continued*)

Table 3. (*Continued*)

Factor	Code item	Number item	Item SEAQQ: In general, at the beginning of this year of study I expected: SPAQQ: In general, at this moment I perceive:	SEAQQ α	SPAQQ α
4. Interpretation of test scores 4 items	V05	25	My scores on tests reflect the extent to which I have mastered the subject.	.88	.76
	V06	26	My scores on various tests on the same topic are comparable.		
	V08	27	I would score the same for a test if different questions or tasks about the same subject were presented to me.		
	R03	28	I would get more or less the same score on a test if I took the test for a second time (supposing my understanding of the subject matter has remained the same).		
5. Authenticity of assessment 5 items	T06	2	Testing and assessment correspond with the activities I will have to perform in my future occupation.	.82	.76
	T07	8	I understand testing and assessment.		
	T13	9	The circumstances in which I am tested or assessed are similar to the working conditions of my future profession.		
	T14	10	Testing and assessment unveil my thinking processes, for instance when I am asked to underpin certain choices.		
	T16	11	I need the competences I require to pass my tests in other (professional) situations as well.		
6. Credibility of assessment 6 items	V01	17	I agree with the manner in which I am examined.	.89	.81
	V04	35	The teachers use the results of the tests and assessments to adjust the teaching.		
	R01	22	Judgements are made independently of the persons who rate me.		
	R02	23	Assessments are made independently of the situations I am assessed in.		
	T03	42	I trust testing and assessment in my educational programme to be of good quality.		
	T04	43	I get actively involved in testing and assessment in my educational programme.		

Note: The codes of the items V (validity), T (transparency), R (reliability), or O (other) refer to the four subscales based on a literature review. Two items were removed because they do not load: T15 (“I also self-assess during my courses. For instance, I assess my report before I hand it in”), no loading in SPAQQ; V09 (“The same competences are assessed several times in various tests”), no loading in SEAQQ. Two items were removed due to poor reliability: T12 (“Identical tests are available for me to practice for the real test”) and T19 (“Testing and assessment are also suitable for students who are impaired in one way or the other, for instance deaf or dyslectic students”). The items were originally presented in Dutch, and they have been translated into English.

4. Discussion

It is necessary to include students’ unique perspectives when attaining assessment quality because students can influence that quality. Therefore, this study aimed to construct two evaluation tools, the SEAQQ and the SPAQQ, to measure students’ expectations and perceptions of assessment quality. The validation process, consisting of PAF, confirmed that a six-factor structure provides an acceptable conceptual basis for capturing students’ expectations and perceptions of

Table 4. Examples of the students' comments on the open questions related to the items of the questionnaires

Code item	Item	Students' comments SEAQQ: At the beginning of this year of study I expected: SPAQQ: At this moment I perceive:
T07	I understand testing and assessment.	SEAQQ: The learning content and assessment instruments should be clear.
		SPAQQ: Now it is clearer to me. The assessment instruments help me focus.
V02	The tests correspond with the learning targets.	SEAQQ: That the test items are literally retrieved from the study books.
		SPAQQ: There are assessments that do not correspond with the learning targets.
O01	The difficulty of testing and assessment concur with the level of my education.	SEAQQ: I expected the assessments in higher education to be more difficult than in secondary education.
		SPAQQ: Good, the assessments are on my level.
V11	The tests and assessments are organised well.	SEAQQ: Assessment should be well organised. At short notice, grades should be available and students should be allowed to inspect the examinations.
		SPAQQ: The assessments are planned in the evening, this disturbs my concentration.
R06	All tests feature correct language.	SEAQQ: Consistent language, no misleading test items.
		SPAQQ: Often there are spelling mistakes in the knowledge tests.

Note: The quotes were translated from Dutch to English.

assessment quality: (1) effects of assessment on learning, (2) fairness of assessment, (3) conditions of assessment, (4) interpretation of test scores, (5) authenticity of assessment, and (6) credibility of assessment.

The six-factor structure can be interpreted as follows. The first factor, "effects of assessment on learning", represents the influence of assessment on students' learning processes and their progress (Assessment reform group, 2002; Hattie, 2009). It contains items, such as self-regulation, feedback, and motivation. The second factor, "fairness of assessment", refers to whether the requirements for successfully taking the assessment are reasonable and feasible; for example, the correspondence between the tests and the learning goals. The third factor, "conditions of assessment", contains circumstances that impact students but that they cannot control, such as test organisation, teacher professionalism, and test construction. The fourth factor, "interpretation of test scores", refers to the meaning of the students' test scores (Wools, Eggen, & Sanders, 2010), such as whether or not the scores reflect the students' actual mastery of the subject. The fifth factor, "authenticity of assessment", represents the alignment of testing and assessment with professional life (Gulikers, Bastiaens, & Kirschner, 2004), such as the similarity of testing conditions to the conditions students will encounter in their future jobs. The sixth factor, "credibility of assessment", refers to the students' belief in assessment; it contains items about trust and involvement. In the final structure of the SEAQQ and the SPAQQ, each factor showed a sufficient degree of reliability (α = .76 to α = .94) (Brace et al., 2016).

While this factor structure can be meaningfully interpreted, it does not match the dimensions found in the literature: validity, transparency, and reliability (Gerritsen-van Leeuwenkamp et al., 2017). This is in line with previous research, which found that there are differences between students' perspectives and the perspectives of teachers (Meyer et al., 2010; Van de Watering & Van de Rijt, 2006), and between students' and assessment specialists' concepts of assessment quality (Sambell et al., 1997). This implies that these questionnaires cannot be used directly by stakeholders who are not students. Further research is required to collect validity evidence to justify the use of the questionnaires with other stakeholders. A comparison of factor structures could identify differences and similarities between the conceptualisations of assessment quality by different groups of stakeholders. This would establish a better consensus on the conceptualisation of assessment quality.

One limitation of PAF, the method used in this study, is that the obtained conclusions are restricted to this sample (Field, 2013). In order to generalise the results, further research should focus on cross-validation (Field, 2013), using several samples of students in other courses and at universities in different countries. For example, previous research found differences in first-year students' perceptions of their own capabilities due to differences in their education, age, and gender (Lizzio & Wilson, 2004). Furthermore, students (primarily in higher health care education programmes) at two universities of applied sciences in the Netherlands were involved in the present research study, and more females than males were represented in the sample.

Further evidence of validity should be collected to support the interpretations and inferences for other uses (AERA, APA, & NCME, 2014). Samples of students from other disciplines and from other countries are necessary to generalise the findings of the present study, because, for example, cultural differences could lead to varying interpretations of the items (Sperber, 2004). Furthermore, the students in the present study focused on assessment quality in general; further validation could focus on using the questionnaires in more specified contexts, i.e. by evaluating the quality of classroom assessment. In this study, the students responded to the SEAQQ and the SPAQQ at the same time. Interaction between those questionnaires may have influenced the outcomes; therefore, further research on administering these two questionnaires separately is needed.

In summary, this study developed two questionnaires, SEAQQ and SPAQQ, to measure students' expectations and perceptions of assessment quality. Educational organisations can use these questionnaires to evaluate assessment quality from a students' perspective. The results of the evaluation can be compared with the perspectives of other stakeholders, and differences and contractions can be discussed while assuring, monitoring, and improving assessment quality. In addition to this study's practical relevance, it provides some insight into students' perspectives on assessment quality, thereby filling a gap in the current scientific research. The results of this study provide a foundation for further research among stakeholders to reach a consensus about what assessment quality is in order to improve and guarantee overall assessment quality in higher education.

Funding
The authors received no direct funding for this research.

Author details
Karin J. Gerritsen-van Leeuwenkamp[1,2]
E-mail: karin.gerritsen-vanleeuwenkamp@ou.nl
ORCID ID: http://orcid.org/0000-0001-8749-9921
Desirée Joosten-ten Brinke[1,3]
E-mail: desiree.joosten-tenbrinke@ou.nl
ORCID ID: http://orcid.org/0000-0001-6161-7117
Liesbeth Kester[4]
E-mail: l.kester@uu.nl
ORCID ID: http://orcid.org/0000-0003-0482-0391
[1] Welten Institute of the Open University of The Netherlands, Heerlen, The Netherlands.
[2] Quality Assurance Office of Saxion, University of Applied Sciences, Enschede, The Netherlands.
[3] Teacher Training Institute, Fontys University of Applied Sciences, Eindhoven, The Netherlands.
[4] Department of Education & Pedagogy, Utrecht University, Utrecht, The Netherlands.

Citation information
Cite this article as: Developing questionnaires to measure students' expectations and perceptions of assessment quality, Karin J. Gerritsen-van Leeuwenkamp, Desirée Joosten-ten Brinke & Liesbeth Kester, *Cogent Education* (2018), 5: 1464425.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, T., & Rogan, J. M. (2010). Bridging the educational research-teaching practice gap. Tools for evaluating the quality of assessment instruments. *The International Union of Biochemistry and Molecular Biology*, *38*, 51–57. doi:10.1002/bmb.20362

Archer, J., & McCarthy, B. (1988). Personal biases in student assessment. *Educational Research*, *30*, 142–145. doi:10.1080/0013188880300208

Assessment reform group. (2002). *Assessment for learning: 10 principles*. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Afl_principles.pdf

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, *32*, 153–170. doi:10.1016/j.stueduc.2006.04.006

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007a). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review*, *2*, 114–129. doi:10.1016/j.edurev.2007.06.001

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007b). Teachers' opinions on quality criteria for competency assessment programs. *Teaching and Teacher Education*, *23*, 857–867. doi:10.1016/j.tate.2006.04.043

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., … Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review*, *1*, 61–67. doi:10.1016/j.edurev.2006.01.001

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*, 151–167. doi:10.1080/713695728

Brace, N., Kemp, R., & Snelgar, R. (2016). *SPSS for psychologists and everybody else*. New York, NY: Palgrave.

Brown, G. T. L., McInerney, D. M., & Liem, G. A. D. (2009). Student perspectives of assessment, considering what assessment means to learners. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 1–21). Scottsdale, AZ: Information Age Publishing.

Carroll, C., Booth, A., & Cooper, K. (2011). A worked example of "best fit" framework synthesis: A systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, *11*(29), 1–9. doi:10.1186/1471-2288-11-29

Creswell, J. W. (2008). Survey designs. In A. C. Benson & C. Robb (Eds.), *Educational research. Planning, conducting, and evaluating quantitative and qualitative research* (pp. 387–430). Upper Saddle River, NJ: Pearson Education International.

Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology*, *92*, 1169–1176. doi:10.1037/0021-9010.92.4.1169

Dijkstra, J., Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education*, *15*, 379–393. doi:10.1007/s10459-009-9205-z

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *The Metric of Medical Education*, *38*, 1006–1012. doi:10.1111/j.1365-2929.2004.01932.x

Ecclestone, K. (2012). Instrumentalism and achievement: A socio-cultural understanding of tensions in vocational education. In J. Gardner (Ed.), *Assessment and learning* (pp. 140–156). London, UK: Sage Publications.

EHEA. (2014). *Bologna process - European higher education area*. Retrieved from http://www.ehea.info/

ENQA. (2009). *Standards and guidelines for quality assurance in the European higher education area*. Retrieved from http://www.enqa.eu/pubs.lasso

Field, A. (2013). *Discovering statistics using SPSS*. London, UK: Sage Publications.

Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation*, *55*, 94–116. doi:10.1016/j.stueduc.2017.08.001

Gulikers, J., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*, 67–86. doi:10.1007/BF02504676

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence-based assessment quality. *Studies in Educational Evaluation*, *35*, 110–119. doi:10.1016/j.stueduc.2009.05.002

Gulikers, J., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction*, *18*, 172–186. doi:10.1016/j.learninstruc.2007.02.012

Harvey, L., & Green, D. (1993). Defining quality. *Assessment & Evaluation in Higher Education*, *18*, 9–34. doi:10.1080/0260293930180102

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, UK: Routledge.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Klemenčič, M. (2012). Student participation in higher education governance in Europe. *International Higher Education*, *66*, 32–33. Retrieved from https://ejournals.bc.edu/ojs/index.php/ihe/article/view/8590/7722

Könings, K. D., Brand-Gruwel, S., van Merriënboer, J. J. G., & Broers, N. J. (2008). Does a new learning environment come up to students' expectations? A longitudinal study. *Journal of Educational Psychology*, *100*, 535–548. doi:10.1037/0022-0663.100.3.535

Levin, B. (1998). The educational requirement for democracy. *Curriculum Inquiry*, *28*, 57–79. doi:10.1111/0362-6784.00075

Levin, B. (2000). Putting students at the centre in education reform. *Journal of Educational Change*, *1*, 155–172. doi:10.1023/A:1010024225888

Linn, R. L., Bakker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21. doi:10.3102/0013189X020008015

Lizzio, A., & Wilson, K. (2004). First-year students' perception of capability. *Studies in Higher Education*, *29*, 110–128. doi:10.1080/1234567032000164903

McMillan, J. H. (2007). Formative classroom assessment: The key to improving student achievement. In J. H. McMillan (Ed.), *Formative classroom assessment. Theory into practice*. New York, NY: Teachers College Press.

Meyer, L. H., Davidson, S., McKenzie, L., Rees, M., Anderson, H., Fletcher, R., & Johnston, P. M. (2010). An investigation of tertiary assessment policy and practice: Alignment and contradictions. *Higher Education Quarterly*, *64*, 331–350. doi:10.1111/j.1468-2273.2010.00459.x

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12. doi:10.3102/0013189X023002005

Moss, P. A., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement: Interdisciplinary Research and Perspectives*, *3*(2), 63–83. doi:10.1207/s15366359mea0302_1

O'Donovan, B. (2016). How student beliefs about knowledge and knowing influence their satisfaction with assessment and feedback. *Higher Education*, 1–17. doi:10.1007/s10734-016-0068-y

Roese, N. J., & Sherman, J. W. (2007). Expectancy. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 91–115). New York, NY: Guilford Press.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*, 349–371. doi:10.1016/S0191-491X(97)86215-3

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education*, *38*, 974–979. doi:10.1111/j.1365-2929.2004.01916.x

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. doi:10.3102/0013189X029007004

Sperber, A. D. (2004). Translation and validation of study instruments for cross-cultural research. *Gastroenterology*, *126*, S124–S128. doi:10.1053/j.gastro.2003.10.016

Stobart, G. (2008). *Testing times. The uses and abuses of assessment*. Oxon, NY: Routledge.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, *30*, 331–347. doi:10.1080/02602930500099102

Svensson, G., & Wood, G. (2007). Are university students really customers? When illusion may lead to delusion for all! *International Journal of Educational Management*, *21*, 17–28. doi:10.1108/09513540710716795

Van de Watering, G., & Van de Rijt, J. (2006). Teachers' and students' perceptions of assessment: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*, 133–147. doi:10.1016/j.edurev.2006.05.001

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & Van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, *34*, 205–214. doi:10.3109/0142159X.2012.652239

Wise, S. L., & Cotten, M. R. (2009). Test-taking effort and score validity. The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187–205). Scottsdale, AZ: Information Age Publishing.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *Cadmo*, *1*, 63–82. doi:10.3280/CAD2010-001007

Zilberberg, A., Brown, A. R., Harmes, J. C., & Anderson, R. D. (2009). How can we increase student motivation during low-stakes testing? In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 255–277). Scottsdale, AZ: Information Age Publishing.

Zimbardo, P. G., Weber, A. L., & Johnson, R. L. (2009). Sensation and perception. In S. Frail, J. Swasey, D. Hanlon, & A. Pickard (Eds.), *Psychology core concepts* (pp. 287–333). Boston, MA: Pearson Education.