Studies in Educational Evaluation 55 (2017) 94–116

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Assessment quality in tertiary education: An integrative literature review

Karin J. Gerritsen-van Leeuwenkamp[a,b,*], Desirée Joosten-ten Brinke[a,c], Liesbeth Kester[d]

[a] Open University of The Netherlands, PO Box 2960, 6401 DL Heerlen, The Netherlands
[b] Saxion, University of Applied Sciences, PO Box 70.000, 7500 KB Enschede, The Netherlands
[c] Fontys University of Applied Sciences, PO Box 347, 5600 AH Eindhoven, The Netherlands
[d] Utrecht University, PO Box 80125, 3508 TC Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Assessment quality
Higher education
Vocational education
Tertiary education
Text analysis
Literature review

ABSTRACT

In tertiary education, inferior assessment quality is a problem that has serious consequences for students, teachers, government, and society. A lack of a clear and overarching conceptualisation of assessment quality can cause difficulties in guaranteeing assessment quality. Thus, the aim of this study is to conceptualise assessment quality in tertiary education by providing an overview of the assessment quality criteria, their influences, the evaluation of the assessment quality criteria, and the perspectives that should be considered when evaluating assessment quality. This study aggregated 78 peer-reviewed journal articles in a framework using MAXQDA, and a text analysis was performed using Leximancer. The results identified validity, transparency, and reliability as assessment quality criteria; standardisation, stakeholders, clarity, and construct irrelevant variance as influences on the assessment quality criteria; validation and statistical data analyses to evaluate assessment quality; and students, staff, government, and experts as perspectives that should be considered when evaluating assessment quality. These insights are important for teachers, educational advisors, and managers who can use this information to determine what assessment quality means for their educational organisation and what they should consider when guaranteeing assessment quality. Moreover, the study provides researchers with insight into the current state of scientific evidence.

1. Introduction

Assessment quality includes the quality of all aspects of assessment practices, such as test items, tasks, assessments, tests, the process of assessing, a programme of assessments in a course or a curriculum and the procedures, policies, and administration of the assessment process. Inferior assessment quality is a problem with serious consequences at all levels of education. It impacts the suitability, accuracy, and credibility of information about students' performance and progress that is collected for the purposes of learning, selection and certification, and accountability (Anderson & Rogan, 2010; Hambleton & Murphy, 1992; Stobart, 2008). However, in tertiary education, the consequences of inferior assessment quality have a significant impact on students and their functioning in society. The information obtained from the assessment will be used to make important decisions about whether a student meets the demands that society imposes on graduates, namely, undertaking occupations, leveraging advanced knowledge, thinking critically, engaging in lifelong learning, and contributing to future innovation (Ministry of Science Technology and Innovation, 2005).

1.1. Consequences of inferior assessment quality

In terms of the learning purpose of assessment, inferior assessment quality can prevent students and teachers from determining the extent to which students have progressed in their learning process, what learning objectives have already been achieved or need to be achieved, and how these objectives can be accomplished (Assessment reform group, 2002; Hattie, 2009). Vaguely formulated learning goals do not provide students with information about the direction of their learning process (Hattie & Timperley, 2007). Furthermore, even when clear learning goals are available, assessment tasks do not always address them. Consequently, students' attention may stray away from learning goals, which may hinder effective learning (Boud and Associates, 2010). During the assessment process, students receive feedback from teachers or peers. However, this feedback can be ineffective when it fails to focus on performance related to learning goals (e.g. when it focuses on students' performance related to others), since such feedback does not provide information on how students can improve their learning (Black & Wiliam, 1998). All these examples obstruct learning, which can lower achievement and lead to a failure to complete

* Corresponding author at: Saxion University of Applied Sciences, PO Box 70.000, 7500 KB Enschede, The Netherlands.
E-mail addresses: karin.gerritsen-vanleeuwenkamp@ou.nl (K.J. Gerritsen-van Leeuwenkamp), desiree.joosten-tenbrinke@ou.nl (D. Joosten-ten Brinke), l.kester@uu.nl (L. Kester).

http://dx.doi.org/10.1016/j.stueduc.2017.08.001
Received 15 February 2017; Received in revised form 28 July 2017; Accepted 2 August 2017
Available online 22 September 2017
0191-491X/

education at all (Gibbs, 2010; Stiggins, 2007).

Inferior assessment quality also impacts the selection and certification purpose of assessment because it can cause unfair failing and unfair passing. Due to inter-rater differences, students might receive lower or higher grades when judged by different assessors (Reynolds, Livingston, & Willson, 2010). Insufficient assessment quality could eventually allow students to receive diplomas even when they have not reached the required exit level. Such a situation is detrimental to the education system as a whole because, if students cannot meet requirements, standards will be lowered, diplomas will be devaluated, and the accountability of graduates and tertiary education institutions will be diminished (Meyer et al., 2010; QAA, 2014; Van der Vleuten et al., 2012). Thus, public funds invested in tertiary education, as well as students' and parents' payments of tuition fees (Yang & McCall, 2014), may reap lower social and personal returns.

Furthermore, inferior assessment quality has consequences for the accountability purpose of assessment. It can result in an inaccurate impression of students' performances, based on what individuals or groups are held accountable for. When students are taught to the test, they perform better on national exams, but they may not actually learn more (Stobart, 2008) or be adequately prepared for their future professions. In contrast, when students perform worse on tests (e.g. due to the complexity of the language used), the assessment may be measuring secondary abilities (e.g. reading abilities) instead of their proficiency in the intended subject (Stobart, 2008). Such discrepancies can lead to unnecessary quality improvement activities or unjustified rewards or sanctions (e.g. an educational institution may be placed under control).

1.2. Difficulties in guaranteeing assessment quality

For the reasons presented above, educational institutions have an obligation to guarantee assessment quality. This is a difficult task for teachers, educational advisors, managers, and researchers because no clear and overarching conceptualisation of assessment quality is currently available. Books offer descriptions of assessment quality, and current standards, codes, frameworks, and guidelines (e.g. AEA (2012), AERA, APA, and NCME (2014), Gilbert and Maguire (2014), and QAA (2014)) based on expert committees, provide useful instructions for constructing assessments and evaluating assessment quality. Researchers are involved in the construction of these books and guidelines; however, the link between practice and research is not always clear. Books and journal articles provide multiple descriptions of assessment quality, for example of validity (e.g. AERA et al., 2014; Borsboom, Mellenbergh, & Van Heerden, 2004). Misconceptions can occur when stakeholders are using the same concepts with different meanings. Therefore, it is important to clarify how assessment quality is conceptualised and why there are differences in conceptualisations. Furthermore, no review of the literature of assessment quality in tertiary education is available; consequently, there is no overview of the strength of the scientific evidence, where the gaps are in the current research, what relationships exist among the findings and what the theoretical and practical implications are. While standards may provide descriptions of quality criteria as validity, reliability, and fairness, in the literature on tertiary education, additional quality criteria are mentioned, such as authenticity, transparency, and cheat proofness (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Yorke, 2008). It remains unclear if these criteria ought to be considered when evaluating assessment quality or whether relationships exist among all the quality criteria mentioned in the literature. These issues must be resolved to determine the requirements of quality (Harvey & Green, 1993). Several researchers addressed challenges in assessment practices, such as discipline-related marking behaviour and variability in students' performances, which influence assessment quality (Bridges et al., 1999; Van der Vleuten, Norman, & Graaff, 1991), but no overview of these influences is currently available. This also applies to how assessments are evaluated, such as through self-evaluation procedures (Baartman, Prins, Kirschner, & Van der Vleuten, 2011) or item analyses (Anderson & Rogan, 2010). Furthermore, questions remain regarding whether assessment quality is a stable concept, since quality depends on the perspectives of the stakeholders defining it (Harvey & Green, 1993); thus, an overview of these perspectives will provide more insight into the meaning of assessment quality. However, before the research aim and questions are presented, it is necessary to discuss the evolution of assessment over time because, this influences the conceptualisation of assessment quality. Therefore, insight into the historical development of assessment is needed to fully understand differences in the conceptualisation of assessment quality. To address this, the next section will focus on the evolution of assessment based on changes in the demands of society and developments in learning theory and scientific measurement.

1.3. Historical development of assessment

Since the beginning of the 20th century, the main focus in assessment has been on efficiency, following the societal movement to accomplish things with a minimal amount of time and effort (Shepard, 2000). Students were primarily educated in line with their abilities: that is, they learned only those things that were directly required for their profession (Shepard, 2000). Teachers determined the learning goals. Furthermore, students played a passive role in their learning processes and there was little interaction between them and their teachers (Attard, Di Ioio, Geven, & Santa, 2010). Tests measured knowledge, skills, and attitudes in isolation, and they repeatedly determined whether students had achieved their goals and were ready to pursue subsequent goals (Schuwirth & Van der Vleuten, 2011; Shepard, 2000). In line with behaviourism, the process of learning was not addressed; rather, observable behaviour was more important (Driscoll, 2005). Passing a test functioned as a reward for succeeding in one of many learning steps (Shepard, 2000). Tests were designed and conducted in line with psychometrics, which sought to: 'develop interpretations that are generalizable across individuals and contexts and to understand the limits of those generalizations' (Moss, Pullin, Gee, & Haertel, 2005, p. 68). These tests were characterised by standardised procedures of data collection, uniformity, and objectivity, since only standardised tests were able to legitimise the comparison of results at different times and in different places (Moss et al., 2005). Experts measured the technical quality of tests primarily based on psychometric quality criteria (Linn, Bakker, & Dunbar, 1991).

Societal demands about what students should be capable of accomplishing after graduation change over time due to economic and social-cultural developments. The launch of Sputnik (the first satellite) in 1957 was one of the turning points in educational reform in the United States (US). Americans were taken by surprise when the Soviets were able to launch a satellite before they could. Consequently, they paid more attention to the level of educational standards and devoted greater focus to inquiry and problem solving skills (Bybee, 1997). This example illustrates how, although knowledge remained important, graduates of tertiary education institutes began to be expected to be capable of anticipating and adapting to changes in their future work environment of lifelong learning (Baartman et al., 2006; Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007a; Boud & Falchikov, 2006; Boud, 2000). Such objectives require an integration of knowledge, skills, and attitudes, as well as the application of these so-called competencies in different authentic situations (Baartman et al., 2006; Baartman et al., 2007a; Van Merriënboer & Kirschner, 2007). Since assessment has a backwash effect on learning (Boud and Associates, 2010; Watkins, Dahlin, & Ekholm, 2005), researchers have argued for the importance of alignments between learning and assessment (Biggs, 1996; Cohen, 1987).

In addition to the evolving requirements of society, insights into how students learn have continued to increase; this improved knowledge has also affected assessment. The focus shifted from observable

behaviour to how students learn. In line with cognitivism and (social) constructivism, the ways in which information is processed, stored and (re)structured, how students regulate their own learning process, how they apply knowledge in new situations (transfer), and the influence of culture and the social environment on learning, all became increasingly important (Driscoll, 2005; Shepard, 2000). Students engage in more interactions with their teachers and fellow students, they learn actively and they assume the central role in their learning (Attard et al., 2010). It was no longer sufficient for tests to determine whether students achieved goals and were ready to pursue subsequent goals. Assessment was expected to provide insight into students' positions in their learning processes, what learning objectives needed to be achieved, and how these objectives could be met. In addition to the use of standardised tests, assessment began to focus on authentic problem solving assessment tasks, the active role of students, the integration of assessment and learning, and the use of multiple flexible measurements (Birenbaum et al., 2006; Shepard, 2000). An illustrative example of this is a performance assessment, which gathers information on students' performances of authentic tasks in real-life physical and social contexts (Gulikers, Bastiaens, & Kirschner, 2004).

In the first half of the 20th century, the main focus was on certifying; this was also called the summative purpose of tests. In the 1960s, a distinction was made between formative and summative evaluation. Scriven (1967) and Bloom, Hastings, and Madaus (1971) used the terms formative and summative to differentiate between the two roles that evaluation may play in education (Black & William, 2003). An evaluation is formative if it is used in the development or improvement of some educational process. An evaluation is summative if it is used in decision-making concerning the end results of an educational process. Where Scriven (1967) focused mainly on the improvement of courses, according to Bloom et al. (1971), formative evaluation informs teachers about students' learning. Sadler (1989) added that students could also use this information to improve their performance; for this purpose, he used the term formative assessment instead of formative evaluation. Formative assessment, such as classroom questioning and feedback, provides information that can be used to alter teaching and learning (Black & Wiliam, 1998; Knight, 2001). Some researchers refer to formative assessment as an assessment instrument, while others refer to it as the process through which an assessment, regardless of its purpose, yields information that can be used by students and teachers to improve teaching and learning (Bennett, 2011). As a process, formative assessment is also called assessment *for* learning (Martinez & Lipson, 1989), this contrasts with assessment *of* learning, which refers to summative assessment.

Assessments are integrated into consciously composed assessment programmes to gather information from different sources about the acquisition of competences and to support students' learning (Baartman et al., 2006; Gibbs & Simpson, 2004; Van der Vleuten & Schuwirth, 2005). Furthermore, one reason why higher education institutions choose these assessment programmes is related to the integrated complex nature of assessment, which makes it much more difficult to achieve quality solely based on individual assessments. Bloxham, den-Outer, Hudson, and Price (2016) and Bloxham, Boyd, and Orr (2011) affirmed this by noting that when assessors evaluate complex task performance in higher education it can be challening for them to interpret the criteria, to arrive at a reliable judgment, and to achieve inter-rater agreement. In addition, researchers have called for assessment to be judged based on its purpose (Baartman et al., 2006; Van der Vleuten et al., 2012) and its consequences for learning and students (Boud, 2000). It is no longer sufficient to judge assessment solely from a psychometric viewpoint. Moss et al. (2005) stated: 'Whereas in psychometrics, individuals and context are treated as distinct . . . from a sociocultural perspective, individuals and contexts are mutually constitutive such that individuals assume different "identities" . . . in different social contexts . . .' (p. 69). Therefore, in addition to traditional quality criteria, such as objectivity, more qualitative criteria, such as authenticity (Gulikers et al., 2004), educational consequences (Baartman et al., 2006) and cognitive complexity (Linn et al., 1991), are used to evaluate the quality of assessment.

1.4. Research aim and questions

To summarise, inferior assessment quality might have negative consequences for the purposes of learning, selection and certification, and accountability. Therefore, educational institutions have an obligation to guarantee assessment quality. However, this is a difficult task because no clear and overarching conceptualisation of assessment quality is currently available; moreover, this is influenced by historical developments. For all these reasons, the present study aims to conceptualise assessment quality in tertiary education. Summarising and organising the results of the extant articles should give teachers, managers, advisors of educational organisations, and researchers an understanding of the evolution of the concept of assessment quality over time. This review will highlight patterns in the data and make recommendations for further research. The four research questions of this review study are:

1. Which assessment quality criteria appear in the literature?
2. What are influences on assessment quality criteria?
3. How are assessment quality criteria evaluated?
4. Which perspectives are identified in the evaluation of assessment quality?

2. Method

2.1. Search method

The research database EBSCO was systematically searched. The search path presented in Fig. 1 shows the search terms that were used related to assessment and quality and the specifications of the context. Only articles focusing on tertiary education, written in English, and published in peer-reviewed journals from 1998 to 2014, were used. This resulted in 396 hits after removal of duplicates.

To be included in this review, a study had to address and operationalise assessment quality in tertiary education, so all the abstracts were read to verify this. A total of 33 articles remained. Using the snowball method (Fig. 1), in which the reference list of each publication was screened for additional articles related to assessment quality, a total of 91 articles were found. The abstracts of these 91 articles were read to determine whether the articles focused on tertiary education and operationalised assessment quality, and were published in peer-reviewed journals. Following the application of these selection criteria, a total of 45 articles remained.

A total of 78 journal articles were further analysed. Of these articles, 41 were opinion articles, 26 were non-experimental articles, 1 was an experimental article, 5 were opinion/review articles, 1 was a non-experimental/review article, and 4 were review articles. The articles originated from Australia, Israel, Malaysia, New Zealand, South Africa, the Netherlands, the United Kingdom (UK) and the US. The methodology is described in detail in the Appendix A. Table A1 presents information about each of the journal articles included in this review study.

2.2. Synthesis of the studies

Since the focus of this integrative literature review is an overview of research on assessment quality, the qualitative framework synthesis approach was used (Carroll, Booth, & Cooper, 2011; Dixon-Woods, 2011; Gough, Thomas, & Oliver, 2012). The information in each journal article was systematically and explicitly aggregated in a framework based on the research questions (Dixon-Woods, 2011). The first author coded all 78 journal articles using the coding framework and the data-

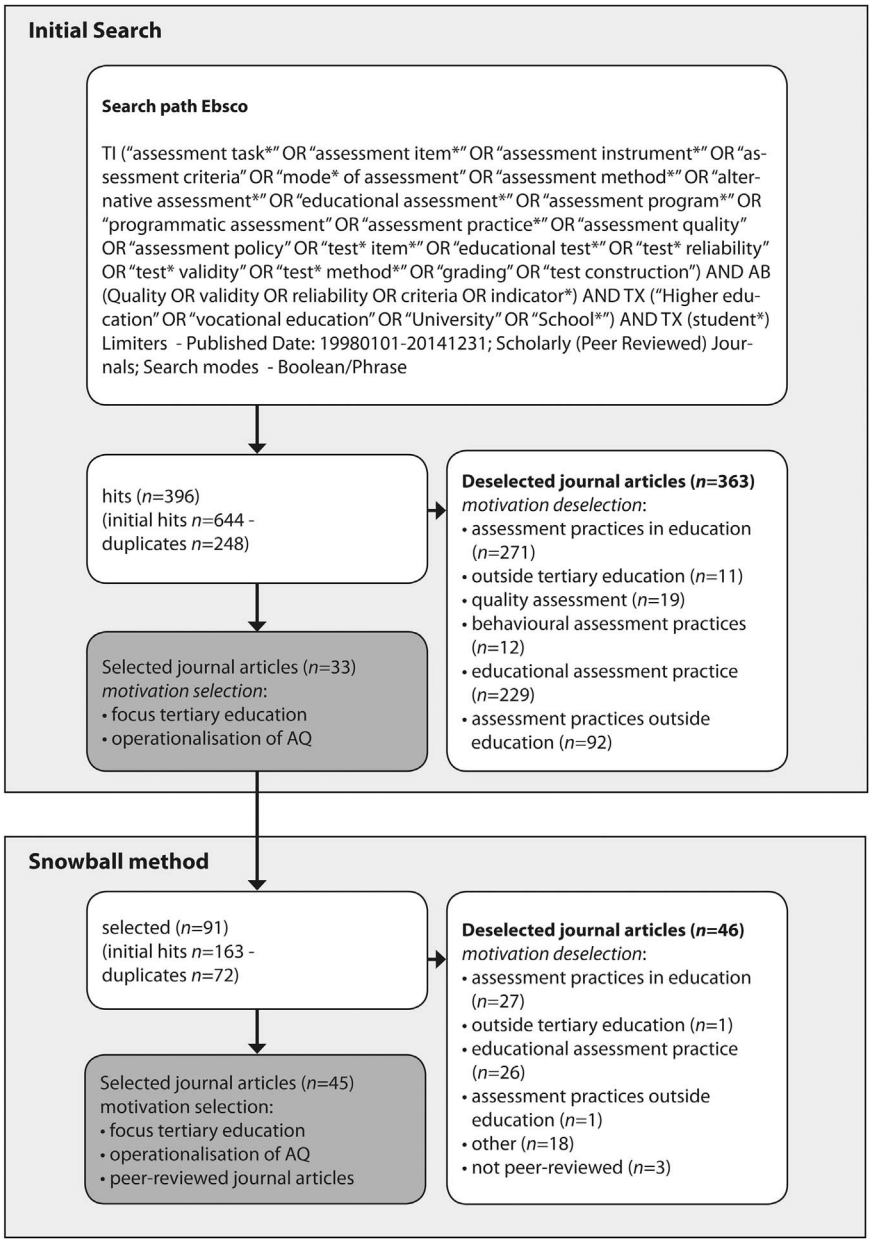

Fig. 1. Search path of the initial search in EBSCO and the process of the snowball method in which the reference list of each of the selected journal articles was screened for useful articles. (De)selection is based on reading the abstract. AQ: assessment quality; TI: title; AB: abstract; TX: text.

analysis software MAXQDA version 11. The second author evaluated the coded segments.

Leximancer 3 was used to execute a conceptual analysis of all the selected text segments, to determine the presence and frequency of concepts in the text and, as a relational analysis, to determine how the identified concepts were related to one another (Leximancer, 2008). Leximancer highlighted the relationships in the data and clustered the concepts into themes. Leximancer provided maps based on the text segments per code. This is depicted in Figs. 2 and 3. As seen in Fig. 2, the assessment quality criteria (grey dots) are clustered into three themes. The size of the themes was set to 50%, which means that 50% of all the criteria that were most frequently connected were clustered into one theme. The other criteria were distributed over the other themes. Fig. 3 depicts the assessment quality criteria within the theme validity and their interrelationships. A comparable process was followed for research questions two, three, and four. Since text analysis was used as a data reduction technique, not all the criteria, influences, evaluations, and perspectives will be discussed in this paper to enhance the comprehensibility and readability. The figures are depicted in detail on the website: www.assessmentquality.com. After data reduction with Leximancer, five of the selected articles did not provide content that was relevant to the research questions (Dennis, 2007; Ediger, 2001; Malouff, 2008; Maxwell, 2012; Segers, Dochy, & De Corte, 1999).

3. Results

3.1. Assessment quality criteria

This section addresses the first research question: Which assessment quality criteria appear in the literature? Based on the text analysis, 98 assessment quality criteria were grouped into three themes: validity, transparency, and reliability (Fig. 2).

3.1.1. Validity

Validity is the first theme of the assessment quality criteria (Fig. 3). It is the most frequently mentioned quality criterion (n = 446), and it is

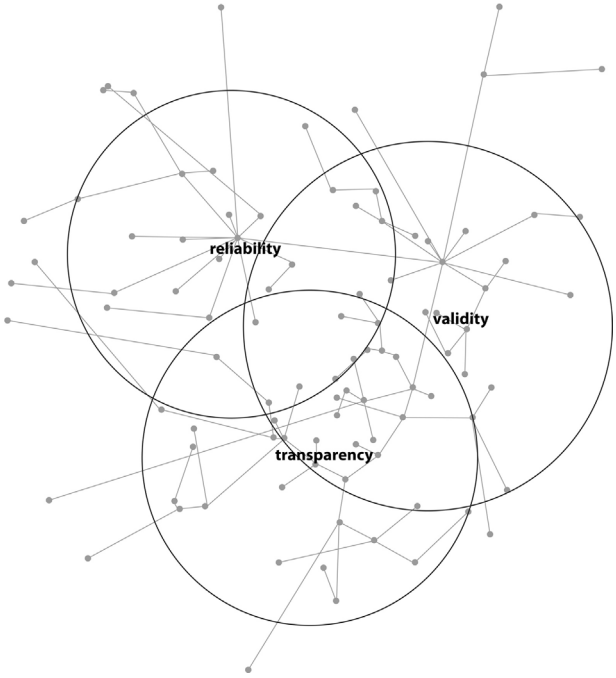

Fig. 2. The themes of assessment quality criteria. The circles depict the themes; the grey dots depict the assessment quality criteria that are grouped together within the themes. The lines depict their interrelations.

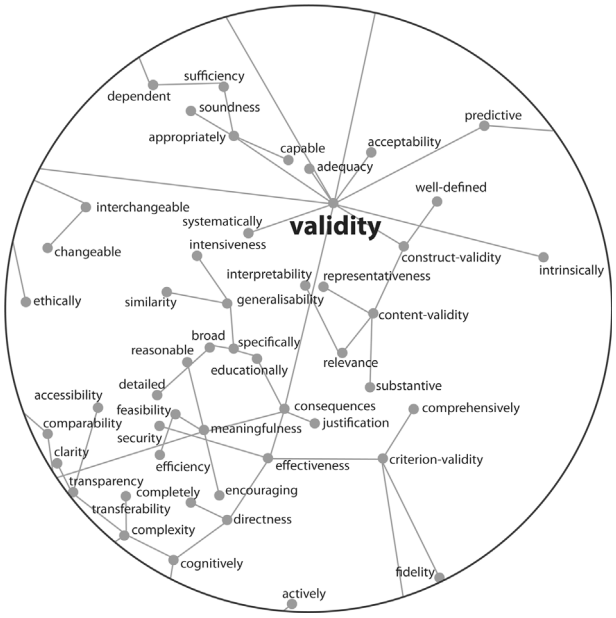

Fig. 3. An overview of the quality criteria within the theme validity.

also the one that is most connected to the other quality criteria. This section presents different viewpoints on validity.

Some authors consider validity to be the characteristic of a test that determines if a test measures what it purports to measure (Anderson & Rogan, 2010; Benett, 1993; Borsboom et al., 2004; Colliver, Conlee, & Verhulst, 2012; Ebel, 1983; Schuwirth & Van der Vleuten, 2003; Van de Watering & Van de Rijt, 2006; Van der Vleuten & Schuwirth, 2005). This definition requires that what is being measured exists and is observable (Borsboom et al., 2004; Ebel, 1983). Other researchers have labelled that definition of validity as content validity, and they have also distinguished other types of validity, such as face validity, consequential validity, construct validity, and criterion validity (including the subtypes: concurrent and predictive validity) (Benett, 1993; Colliver et al., 2012; Cronbach & Meehl, 1955; Harnisch & Mabry, 1993; Martin, 1997; Sambell, McDowell, & Brown, 1997; Van der Vleuten, 1996).

Messick (1995) has argued for using construct validity as the unifying concept of validity. He stated that it is undesirable to distinguish among different types of validity since doing so results in an incomplete perspective of validity. Construct validity, then, addresses 'both score meaning and social values in test interpretation and test use' (Messick, 1995, p. 741). An advantage of this approach is that attention is given to elements of the process that may not be intended or foreseen by the assessment designers (Linn et al., 1991). Since assessment has serious intended and unintended consequences for learning and teaching, a system approach to validity is required (Frederiksen, 1989; Messick, 1995; Sambell et al., 1997). Instead of answering the earlier described question: 'Does a test measure what it purports to measure?' (Shepard, 1993, pp. 409–410), the central question appears to be: 'Does the test do what it claims to do?' (Shepard, 1993, p. 444). In this case, validity is no longer considered to be a characteristic of the test (Shepard, 1993); rather, it is, a judgment of the degree to which the evidence and the arguments support the interpretation and inferences based on test scores, including test use (Kane, 1992, 2001; Linn et al., 1991; Maclellan, 2004; Messick, 1995; Shepard, 1993).

In the use of (competence based) assessments, some researchers call for a change in the conceptualisation of validity or emphasise the importance of new quality criteria to give substance to the validity of competence assessments (Baartman et al., 2011; Harnisch & Mabry, 1993; Linn et al., 1991; Wools, Eggen, & Sanders, 2010). Martin (1997) argued:

> However, as notions of fitness for purpose change and as assessment of more qualitative areas also are developed, the conceptions of validity and reliability encompassed within the instruments of assessment must also change accordingly . . . Thus, just as the validity of an instrument of assessment concerns its fitness for purpose . . . so the notion of validity selected has also to match this same purpose. (p. 338)

Assessments are more qualitative in nature, and students must prove their competence in a larger variety of professional situations, since results achieved in one context do not guarantee the acquisition of competences in another context (Baartman, Gulikers, & Dijkstra, 2013; Leigh et al., 2007). These multiple measurements are integrated in assessment programmes in which judgments are made about a students' competence development (Baartman et al., 2006). According to Leigh et al. (2007), 'Validity refers to the accumulated evidence about the effectiveness of specific assessment models in measuring competencies. The interpretation of validity is not based on a single statistic but through a combination of all available evidence about the assessment' (p. 464). A range of criteria for assessment quality are proposed for assessment programmes (Baartman et al., 2006; Baartman et al., 2011) and individual assessments (Dierick & Dochy, 2001) in order to substantiate validity, to make validity more applicable, and to avoid so-called container concepts (Baartman et al., 2011; Leigh et al., 2007). Fig. 3 presents the criteria within the theme of validity. The most frequently mentioned criteria in the literature are meaningfulness (n = 117) and educational consequences (n = 76). Meaningfulness refers to the added value of an assessment or assessment programme for stakeholders, including whether the assessment (programme) offers learning experiences. The term, educational consequences, refers to the intended and unintended consequences of an assessment (programme), such as the positive effects on students' learning (Baartman et al., 2006; Baartman et al., 2007a; Linn et al., 1991). The range of assessment quality criteria is still rooted in the original definitions of validity (Baartman et al., 2013; Baartman et al., 2007a; Segers, Dierick, & Dochy, 2001).

3.1.2. Transparency

Transparency is the second theme of the assessment quality criteria

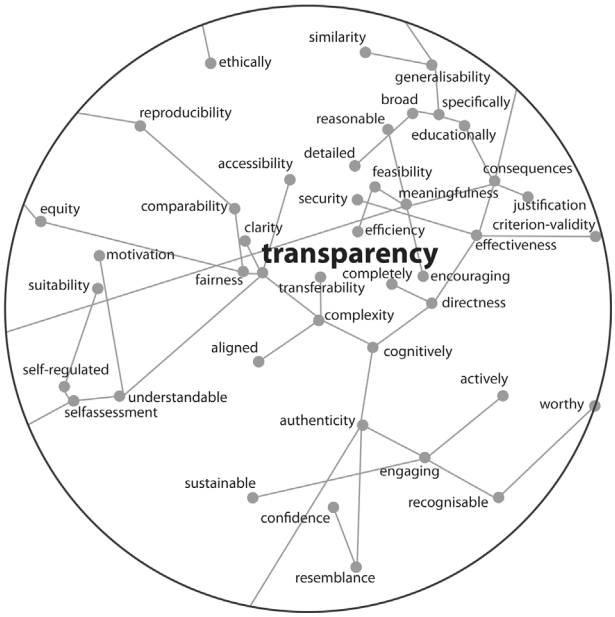

Fig. 4. An overview of the quality criteria within the theme transparency.

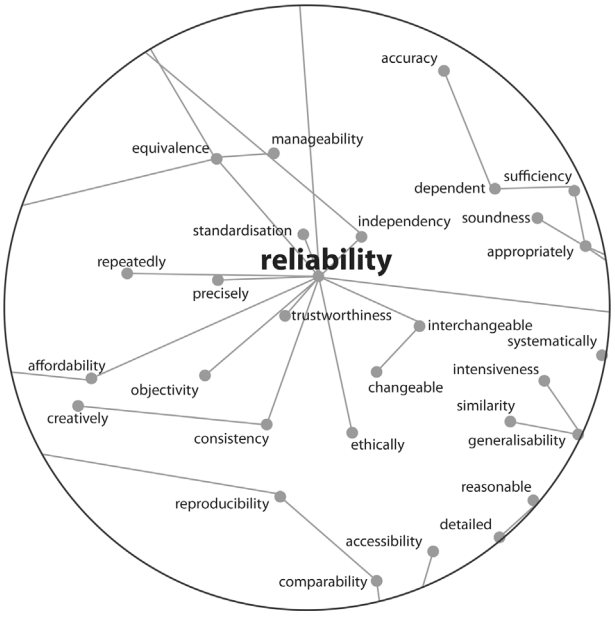

Fig. 5. An overview of the quality criteria within the theme reliability.

(Fig. 4). Although it is mentioned less often than reliability, 74 times versus 187 times, transparency is more related to the other assessment quality criteria than reliability; therefore, it is the second most important theme. Transparency is not explicitly mentioned in relation to testing. It is, however, mentioned in relation to assessment due to the consequences it has on teaching and learning (Frederiksen, 1989; Linn et al., 1991). Linn et al. (1991) stated: 'transparency is considered important because understanding the basis on which performance will be judged facilitates the improvement of performance' (p. 17). Transparency focuses on the assessment process, and it receives more attention in assessments used for learning purposes than in assessments used for the purposes of selection and certification (Ploegh, Tillema, & Segers, 2009). Transparency is essential because it avoids misunderstandings and misinterpretations about what is assessed and how it is assessed (Ploegh et al., 2009).

Transparency applies to items, figures, and used language (Anderson & Rogan, 2010; Verhoeven, Verwijnen, Scherpbier, Schuwirth, & Van der Vleuten, 1999). By using a sample assessment, students know what is expected of them during the official assessment (Anderson & Rogan, 2010). Procedures should adequately document how the assessment proceeds (Dierick & Dochy, 2001; Ploegh et al., 2009), and what is being assessed (Verhoeven et al., 1999). Transparent score reports and feedback are required (Birenbaum, 2007). Goals and criteria should be unambiguously described and explained (Anderson & Rogan, 2010; Gulikers et al., 2004; Woolf, 2004) because an understanding of criteria will stimulate the students' learning (Knight, 2000; Price, Carroll, O'Donovan, & Rust, 2010). Assessors should also be transparent about the restrictions of criteria while assessing students, due to the complexity of the process of assessing and the difficulty of applying the criteria (Woolf, 2004).

Furthermore, transparency refers to assessment programmes (Baartman, Prins, Kirschner, & Van der Vleuten, 2007; Baartman et al., 2011). According to Baartman, Bastiaens, Kirschner, and Van der Vleuten (2007b):

> [Competency Assessment Programs] should be clear and understandable to all stakeholders. Learners should know the scoring criteria, who the assessors are and what the purpose of the assessment is. External controlling agencies should be able to get a clear picture of the way in which a [competency assessment program] is developed and carried out. (p. 860)

3.1.3. Reliability

Reliability is the third theme of assessment quality criteria (Fig. 5). In the text segments retrieved from the journal articles reliability is mentioned 187 times, but it is less connected to the other quality criteria than validity or transparency.

Reliability is considered to be a characteristic of the test score (Downing, 2004). It is important because it determines the accuracy of pass and fail decisions. Decreased reliability results in false positive or false negative outcomes (Downing, 2004). This section presents viewpoints on reliability.

In relation to standardised testing, reliability refers to the degree to which test scores are consistent, precise, and stable, thus, free from errors of measurement (Benett, 1993; Ebel, 1983; Moss, 1994, 1995). Any variability (inconsistency), for example, in a students' performance or in a variety of tasks, is treated as error variance, and thus as a measurement error. The basic assumption behind this definition is that each construct that is being measured is either stable or is treated as stable (Birenbaum, 2007; Schuwirth & Van der Vleuten, 2011). The focus is on objectivity and standardisation (Van der Vleuten & Schuwirth, 2005; Van der Vleuten, 1996).

When more varied assessments are used for the purposes of learning, selection and certification (e.g. performance assessment, self-assessment, oral exams, and problem tasks), the role of human judgment becomes greater, the sample of cases becomes smaller, and authenticity becomes more important. Consequently, there is less consistency over time, more variation, and less reliable assessment scores (Benett, 1993; Knight, 2000; Van der Vleuten, 1996). Instead of striving for reliability on only individual assessments, some researchers have proposed that it is more important to create reliable decisions by using assessment programmes in which multiple, low-stakes and high-stakes assessments are integrated, a large sample of different professional situations is used, and multiple assessors make decisions about the competency development of students (Baartman et al., 2007a; Baartman et al., 2011; Knight, 2000). As Knight (2000) noted:

> Programmatic thinking cannot deliver the chimera of perfect reliability and validity in the assessment of complex learning but it does make it easier to see how to invest in reliability and to identify where it really matters. The corollary is a freedom to think more creatively about assessment where reliable grades are not needed. (p. 239)

In the use of assessment programmes, reliability is achieved in different ways (Baartman et al., 2007a). Scores and decisions are reproduced (Baartman et al., 2011; Downing, 2004; Schuwirth & Van der Vleuten, 2011; Van der Vleuten & Schuwirth, 2005) by repeated criteria-related judgments in which evidence is collected from different sources (Knight, 2000, 2002a; Leigh et al., 2007) based on an adequate sampling of situations (Tweed & Wilkinson, 2012; Van der Vleuten & Schuwirth, 2005). According to Van der Vleuten and Schuwirth (2005), 'what is new, however, is the recent insight that reliability is not conditional on objectivity and standardization' (p. 311). Schuwirth and Van der Vleuten (2004) stated that reliability indicates the degree of generalisability of test scores; the score on the items in the sample should be indicative of a student's scores in any other relevant sample.

Comparable to validity, a range of assessment quality criteria are proposed for assessment programmes (Baartman et al., 2006; Baartman et al., 2011) and individual assessments (Dierick & Dochy, 2001) to make reliability more applicable in practice and to avoid so-called container concepts (Baartman et al., 2011). In those assessment quality criteria, the roots of the original definitions of reliability are still integrated (Baartman et al., 2013). For example, the quality criterion comparability, which means: '[Competency assessment programmes] should be conducted in a consistent and responsible way' (Baartman, Prins et al., 2007, p. 261), refers to the consistency described in previous definitions of reliability.

As shown in Fig. 2, validity and reliability relate to and intersect with each other (Knight, 2000); based on the text analysis, there is a direct relationship between them. In the conceptualisation of content validity, reliability is required for validity. As Ebel (1983) noted, 'unless the test scores measure reliably what the test user intends to measure, the test scores will not be valid' (p. 7). A test will never be valid if it is not reliable (Anderson & Rogan, 2010). Other researchers have emphasised that validity is a condition for reliability because reliability is jeopardized when validity is lacking (Knight, 2002b). Borsboom et al. (2004) stated that, 'it does seem strange to say that "Test X measures intelligence with a certain precision" but that "The test does not measure intelligence"' (p. 1070). In the conceptualisation of construct validity, reliability is integrated into validity (Leigh et al., 2007; Moss, 1994). Some studies take this further by arguing that there can be validity without reliability (Colliver et al., 2012; Moss, 1994), depending on the definition of reliability. Moss (1994) stated 'I now return to my title, "Can there be validity without reliability?" When reliability is defined as consistency among independent measures intended as interchangeable, the answer is, yes' (p. 10). In that case, the assessment may be valid, but not reliable (Colliver et al., 2012).

3.2. Influences on assessment quality criteria

This section addresses the second research question: What are influences on assessment quality criteria? Based on the text analysis, the 73 influences were grouped into four themes: standardisation, stakeholders, clarity, and construct irrelevant variance (Fig. 6).

3.2.1. Standardisation

Standardisation is the first theme of influences (Fig. 7). Standardisation is the theme that is most frequently mentioned ($n = 74$), and it is most connected to other influences. It refers to the standards, specifications, documentation, procedures, guidelines, checklists, scales, and formats used to enhance assessment quality. This section discusses the influence that standardisation has on the three themes of assessment quality criteria validity, transparency, and reliability.

Validity can be improved by scoring instruments, quality control and administration procedures (Van der Vleuten et al., 2012). Predetermined criteria and standards can be used to decide whether the test measures what it purports to measure (Van der Vleuten, 1996), and whether the items and tasks in the assessment are representative of the required difficulty level (Van de Watering & Van de Rijt, 2006). In contrast, non-compliance with forms of standardisation (e.g. neglected procedures) may weaken the interpretive argument (Kane, 1992), thereby undermining construct validity.

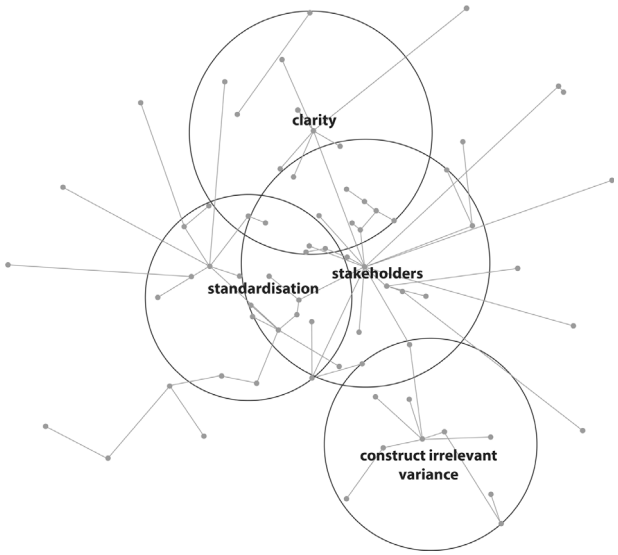

Fig. 6. The themes of the influences on the assessment quality criteria. The grey dots depict the influences that are grouped together within the themes. The lines depict their interrelations.

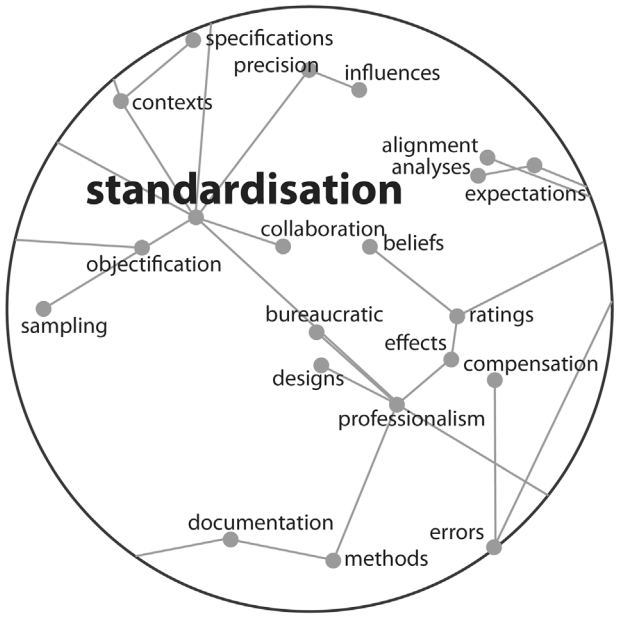

Fig. 7. An overview of the influences within the theme standardisation.

The assessment process becomes more transparent when all the elements of an assessment programme, such as the rights, obligations, rules, and regulations, are documented (Dijkstra, Van der Vleuten, & Schuwirth, 2010). Other forms of standardisation, such as assessment criteria, marking schemes, rating procedures, objectives, and grading sheets or matrices, affect the fairness of assessments, which is an aspect of theme transparency since it gives students information about how their grade is determined (Holmes & Smith, 2003; Stowell, 2004).

Standardisation also influences reliability (Van der Vleuten, 1996). By using a rubric (a descriptive scale), grading can become more consistent (Holmes & Smith, 2003) and, thus, more reliable. Some researchers argued for reliability with less standardisation of methods,

Fig. 8. An overview of the influences within the theme stakeholders.

situations, and performances. Instead, the focus should be on adequate sampling of items, examiners, and time of day to achieve reliability (Van der Vleuten & Schuwirth, 2005; Van der Vleuten et al., 1991; Van der Vleuten, 1996).

3.2.2. Stakeholders

Stakeholders is the second theme of influences (Fig. 8). Although it is mentioned more often than standardisation, 185 times versus 74 times, it is less related to the other influences. Stakeholders are students, the staff of an educational organisation (teachers and examiners), and employers. They influence assessment quality by their use of the assessment, their acceptance of or resistance to it, the extent to which they think the assessment is important, their focus on the assessment instead of on themselves, their involvement in the design of the assessment programmes, and their expertise in the different aspects of the assessment process (Baartman et al., 2007a; Baartman et al., 2011; Boud, 2000; Dijkstra et al., 2010; Price et al., 2010; Schuwirth & Van der Vleuten, 2011; Van der Vleuten et al., 2012). Staff members can influence assessment quality by passing students who are not on the required level by enhancing the rate of a course's success for funding purposes or to avoid students' protests about assessment (Meyer et al., 2010). This section, discusses the influence that stakeholders have on validity, transparency, and reliability.

Stakeholders influence validity in the way they use assessment (Baartman et al., 2007a). When students perceive the consequences of an assessment as being unimportant, or when the assessment does not match their expectations, they might become unmotivated or they might misunderstand components of the assessment. Consequently, their performance could be negatively influenced and the assessment score might not fully reflect their real performance, which, in turn, undermines its validity (Birenbaum, 2007; Downing & Haladyna, 1997; Kane, 1992). Spence-Brown (2001) showed that when students perceive a task as being too difficult or irrelevant, they are less likely to authentically engage with it. This undermines the authenticity of the task, which is a threat to the validity of the assessment (Spence-Brown, 2001). Staff members can jeopardize validity by not complying with assessment procedures or with scoring keys (Kane, 1992). Conversely, validity can be stimulated by involving employers in the assessment construction process because the assessment is more likely to reflect the authentic situation (Gulikers, Biemans, & Mulder, 2009).

Students' perception of fairness, which is an aspect of theme transparency, can be influenced by the consistency of grading procedures, the accuracy of information, disappointing grades, and the presence of biases (Tata, 1999). Separating the role of teacher and examiner, blind marking, or marking each question independently may have a positive impact on fairness (Archer & McCarthy, 1988; Verhoeven et al., 1999). According to students, teachers, and employers, when employers fulfil the role of co-assessor and allow students to perform tasks in a workplace fairness is enhanced (Gulikers et al., 2009). The quality criterion comparability which is an aspect of theme transparency, can be influenced by the employers' uncertainty of their role in the assessment process, the comparability of their judgment to other employers, and the relationship between the student and the employer (Gulikers et al., 2009).

Reliability may be affected by sources of error, such as lower levels of concentration, inattention, guessing, students' feelings of uncertainty and anxiety, and raters' errors (Allen, Reed-Rhoads, Terry, Murphy, & Stone, 2008; Barman, 2011; Downing, 2004; Van de Watering & Van de Rijt, 2006). Increasing a staff's degree of professionalism in the design, evaluation, and invigilation of assessments may influence reliability. Training raters, monitoring the rating process, and aligning the rating with the judgments of other raters may enhance accuracy and consistency, thereby enhancing reliability (Frederiksen, 1989; Leigh et al., 2007). Group sessions with staff members, in which assessment instruments are designed or results are analysed, may lead to a shared vision that can enhance rating consistency (Anderson & Rogan, 2010; Maclellan, 2004). The consistency of the examiners influences the reproducibility of the ratings (Downing, 2004) and, thus, the reliability. Reliability is also strengthened by using more than one rater to judge several parts of an assessment to moderate the bias each rater adds across the assessment (Van der Vleuten, 1996); for example, each essay within a paper is judged by a different rater (Schuwirth & Van der Vleuten, 2004). Van der Vleuten et al. (1991) noted that 'Rater stringency will thus be balanced within examinees (and averaged out), instead of between examinees (favouring some and disadvantaging others)' (p. 116).

3.2.3. Clarity

Clarity is the third theme of influences on assessment quality criteria (n = 39; Fig. 9). A clear description of an assessment programme's goals is essential for assessment quality. In Section 3.2.1, the importance of standardisation via documentation and procedures was

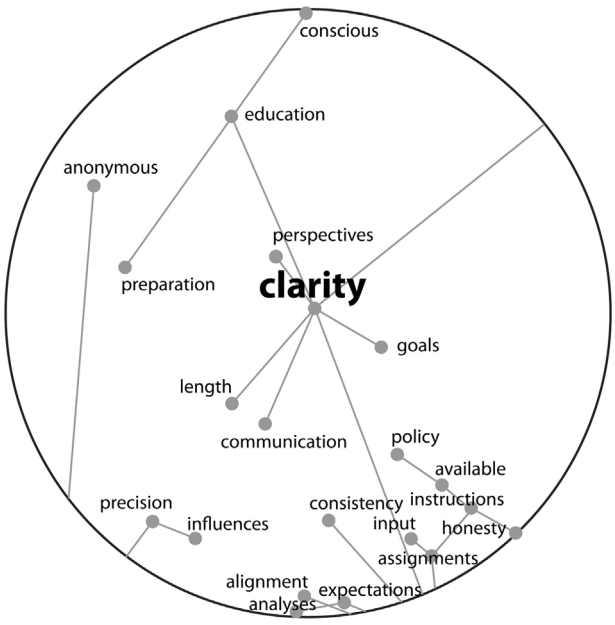

Fig. 9. An overview of the influences within the theme clarity.

described. This section considers the clarity of those documents, and it highlights the influence of clarity on the themes of assessment quality criteria validity, transparency, and reliability.

If students cannot achieve their true level of performance due to unclear tasks, validity decreases (Maclellan, 2004). To ensure the transparency of assessment the goals, rules, and procedures of the assessment programme, the roles and rights of the different stakeholders and the situations where the programme should be used must all be clearly documented (Dijkstra et al., 2010; Gulikers et al., 2009). A clear assessment programme allows its users to determine how the programme influences them, and it helps stakeholders reach agreement (Dijkstra et al., 2010; Gulikers et al., 2009). However, a clear description is not enough. As Van der Vleuten et al. (2012) noted:

> All actors in programmatic assessment should understand what they are doing, why they are doing it and why they are doing it this way. Otherwise they are in danger of losing sight of the true purpose of assessment and will fall back on bureaucratic procedures and meaningless artefacts. (p. 212)

In addition to assessment programmes, individual assessments become more transparent and consistent when stakeholders know what is being assessed (Stowell, 2004). Students should understand the assessment criteria (Baartman et al., 2013). From the perspective of students, the sense of the fairness of assessments, which is an aspect of theme transparency, increases when the goals are clearly described and communicated (Holmes & Smith, 2003). When assessment criteria or items are not clear or comprehensible, reliability can decrease (Anderson & Rogan, 2010; Downing, 2004; Knight, 2002a).

3.2.4. Construct irrelevant variance

Construct irrelevant variance is the fourth theme of influences on assessment quality criteria (n = 24; Fig. 10). It occurs when an assessment measures more than it is supposed to measure. Together with underrepresentation (when a test measures less than it is supposed to measure), construct irrelevant variance is the biggest threat to validity (Messick, 1995; Wools et al., 2010). Messick (1995) distinguished two groups of construct irrelevant variance: construct irrelevant difficulty and construct irrelevant easiness. In construct irrelevant difficulty, for example, the language used in the assessment is too complex, so the assessment measures reading ability instead of what it is supposed to measure (Haladyna, Downing, & Rodriguez, 2002). In construct irrelevant easiness, for example, hints in assessment tasks unintentionally help students perform better (Messick, 1995).

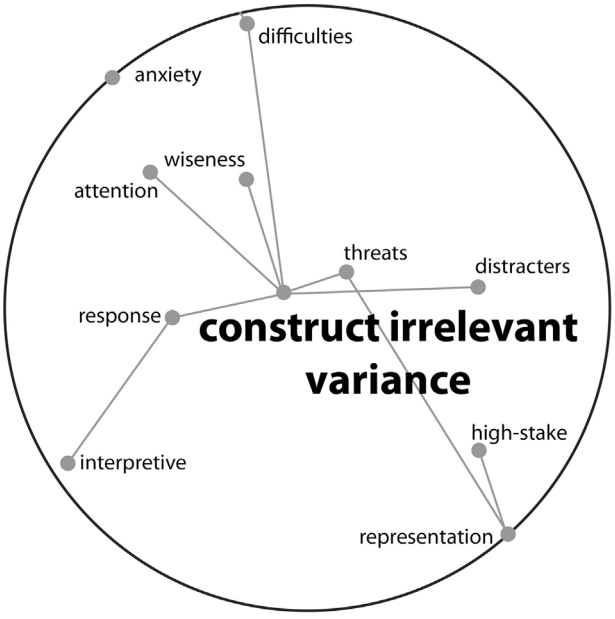

Fig. 10. An overview of the influences within the theme construct irrelevant variance.

3.3. Evaluation of assessment quality criteria

This section addresses the third research question: How are assessment quality criteria evaluated? Based on the text analysis, 22 concepts were grouped into two themes: validation and statistical data analyses (Fig. 11).

3.3.1. Validation

Validation (n = 180) is the first theme associated with the evaluation of the assessment quality criteria. Validation is used to evaluate validity; it is not mentioned in relation to reliability or transparency.

Borsboom et al. (2004) defined validation from the perspective of content validity. They stated: 'In particular, validation is the kind of activity researchers undertake to find out whether a test has the property of validity' (p. 1063). Correlations between test scores and a criterion (i.e. what the test aims to measure) were seen as the best type of evidence for test validity (Ebel, 1983; Shepard, 1993). Whether correlations are the best type of evidence is still under discussion (Borsboom et al., 2004; Ebel, 1983). Ebel (1983) explained that, in most cases, it is impossible to measure a criterion directly:

> What should be used as criterion scores for a test of capability in fifth grade arithmetic? . . . The tests themselves are usually intended to be the best measures of these abilities that can be devised. If better measures were available, the tests would not be needed. (p. 9)

Borsboom et al. (2004) emphasised the need to focus on the causal relationship between the measured attributes and the test scores.

From the perspective of construct validity, validation is defined as the process in which evidence is compiled to support test use and the interpretations of test scores (Downing & Haladyna, 1997; Kane, 2001; Messick, 1995; Shepard, 1993). According to Kane (2001), 'It is the interpretation (including inferences and decisions) that is validated, not the test or the test score' (p. 328). Validation evaluates whether interpretations are correct (Wools et al., 2010), and this is important since it justifies the decisions based on test scores (Kane, 2008).

Several researchers wrote about a (conceptual) framework for validation (Cronbach & Meehl, 1955; Kane, 2001, 2008; Moss, 1995; Shepard, 1993). The argument-based approach is an example of a standardised process of validation that provides such a framework. In this approach, an interpretive argument is based on four phases (Kane, 1992, 2008). These phases are explained by an example from a bachelor's degree in podiatry. In the first phase, the statements or decisions that will be based on the test scores are determined (Kane, 1992, 2008). The test score on an assessment in a real-life situation leads to a decision on proficiency in sole therapy. To take this step, a range of inferences and assumptions must be made. In the second phase, these inferences and assumptions are specified (Kane, 1992, 2008). The sample of tasks in the assessment must include sole therapy of a child, an athlete, and a healthy adult, which are representative of the test domain. In the third phase, potential competing interpretations are identified (Kane, 1992, 2008). Conditions are not equal for each student because students perform the task in different podiatry practices. In the fourth phase, evidence is collected to support the inferences and assumptions and to argue potential counterarguments (Kane, 1992). A reliability study is performed to indicate the way in which the scores are consistent across different situations.

According to Kane (2008), 'The evidence for and against the proposed interpretations and uses provides an overall evaluation of the validity of the claims based on the test scores' (p. 79). Wools et al. (2010) extended Kane's approach with an evaluation phase in which the validation process and the argument are evaluated on three criteria: 'Does the interpretive argument address the correct inferences and assumptions? . . . Are the inferences justified? . . . Is the validity argument

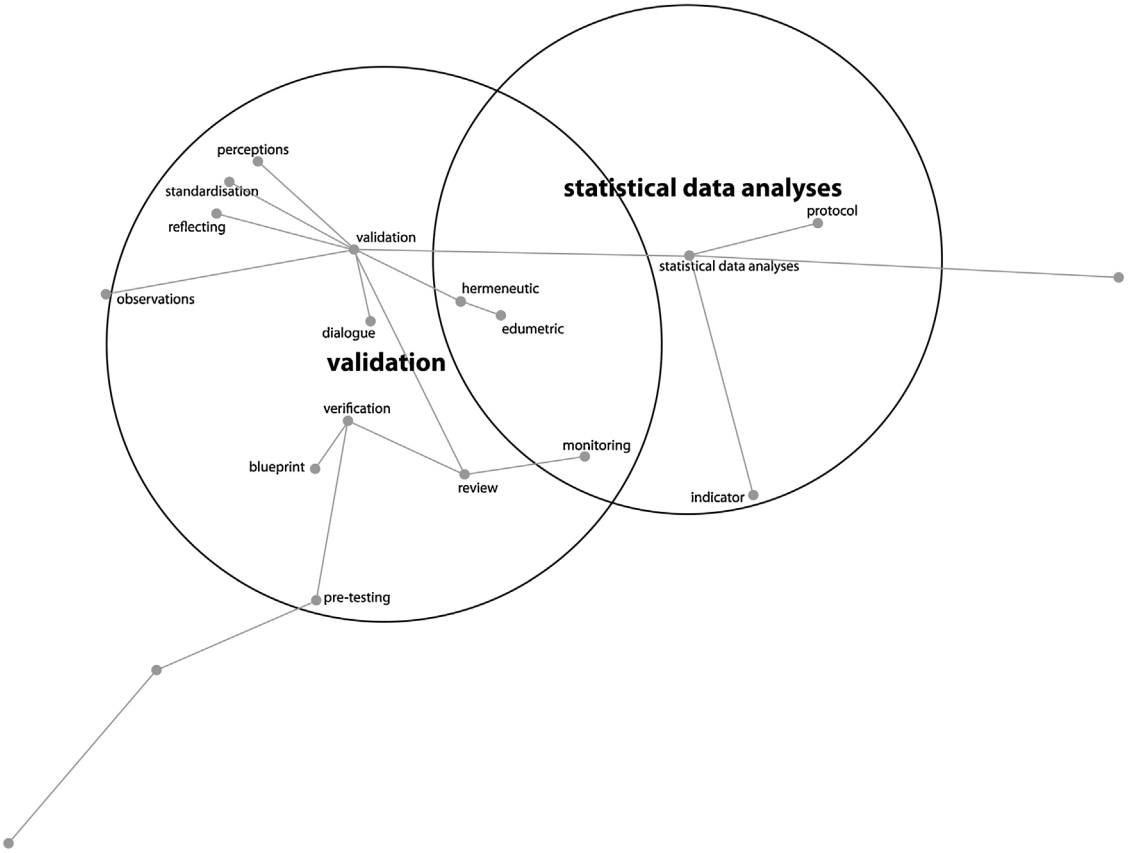

Fig. 11. The themes associated with the evaluation of assessment quality, including the underlying concepts. The grey dots depict the concepts that are grouped together within the themes. The lines depict their interrelations.

as a whole plausible?' (p. 68). Wools et al. (2010) suggested that the argument-based approach might be useful for evaluating assessment quality in general.

How much evidence and what kind of evidence is required depends on the inferences and assumptions (or hypotheses) in the interpretive argument. However, it is important that the evidence consists of different components and addresses the most vulnerable pieces of the argument (Kane, 1992, 2001, 2008; Schuwirth & Van der Vleuten, 2012). Evidence might be empirical, analytical, logically, or rational (Kane, 2001; Messick, 1995; Shepard, 1993; Van der Vleuten, 1996; Wools et al., 2010). It is relevant to collect evidence of the consequences of an assessment, of students' expectations and perceptions of an assessment (Birenbaum, 2007; Messick, 1995), or of justifications for scoring rubrics (Kane, 2001). Evidence of the training, experience, and knowledge of the assessor might also be relevant, since an argument is more persuasive when it is backed by an expert (Schuwirth & Van der Vleuten, 2012). Downing and Haladyna (1997) described an item validity evidence model in which they distinguish 11 types of evidence, the related activities, and the required evidence (i.e. documentation; credentials). The methods used to collect evidence are interviews with stakeholders, expert judgments, and member checking, such as error or reasons analyses (Birenbaum, 2007; Maclellan, 2004; Schuwirth & Van der Vleuten, 2012).

3.3.2. Statistical data analyses

Statistical data analyses (n = 146) is the second theme associated with the evaluation of the assessment quality criteria. This section includes an overview of the evaluation of the quality of test items, in general, as well as the evaluation of the themes of assessment quality criteria, validity and reliability.

Analyses of the test items are useful because they provide information about the features of the test items, such as item difficulty, item discrimination, standard deviation, and distracter frequencies (Anderson & Rogan, 2010; Downing & Haladyna, 1997; McKenna & Bull, 2000; Zakrzewski & Steven, 2003). Item analysis software can generate information to determine the quality of test items (Downing & Haladyna, 1997; Zakrzewski & Steven, 2003).

The correlation between test scores and criterion measures is one type of evidence used to substantiate test validity (Ebel, 1983; Shepard, 1993). Criticisms of this approach were described in the section that addressed validation (Section 3.3.1). The correlation between assessment tasks and representative tasks of the construct domain may be useful evidence to substantiate generalisability (Messick, 1995). Furthermore, methods are available that can be used to explore and interpret the underlying structure of test data, such as confirmatory factor analysis, cluster analysis, and multidimensional scaling analysis (Birenbaum, 2007; Cronbach & Meehl, 1955). According to Birenbaum (2007), 'Such analyses can also be used for justifying the assignment of multiple scores (score profile) rather than a single total score, and the choice of the measurement model for scaling' (p. 35).

Different methods can determine the reliability of a test, based on their characteristics, accuracy, and usefulness to a specific situation (Berk, 1980). The correlation between the scores on two takings of the same test at different times can be used to estimate reliability; this is the so-called test-retest method or the coefficient of stability (Allen et al., 2008; Anderson & Rogan, 2010; Barman, 2011; Cronbach & Meehl, 1955; Downing, 2004; Schuwirth & Van der Vleuten, 2006). Berk (1980) described the calculation of two agreement indices as, 'p0 proportion of individuals consistently classified as masters and nonmasters across (classically) parallel test forms, and k [Kappa], proportion of individuals consistently classified beyond that expected by chance' (p. 327). These can be used to determine the consistency of decisions in the test-retest. The main differences between the two agreement indices are that the p0 index appears to be easier to calculate and the Kappa

statistic corrects for chance agreement (Berk, 1980).

Several researchers have described the impracticality of test-retest methods for educational tests, since the students' knowledge level is not constant during the time in which they take the tests, and the test time and workload are inefficient for both students and staff (Anderson & Rogan, 2010; Barman, 2011; Downing, 2004; Hambleton & Slater, 1997; Schuwirth & Van der Vleuten, 2006). An alternative approach is the split-half method in which the test is split into equal halves and the correlation is determined between those halves. Then, the Spearman-Brown prophecy formula can be used to estimate the reliability of the entire test (Barman, 2011; Hambleton & Slater, 1997; Schuwirth & Van der Vleuten, 2006). One way to determine internal consistency is to automatically split the test into parts, and then compute and mediate the inter-item correlations. Reliability is then expressed by coefficients, including Cronbach's alpha, the Angoff-Feldt coefficient, Kuder and Richardson Formula 20 (KR20) and Kuder and Richardson Formula 21 (KR21), standardised alpha, and maximal reliability (Anderson & Rogan, 2010; Barman, 2011; Berk, 1980; Downing, 2004; Hambleton & Slater, 1997; Schuwirth & Van der Vleuten, 2006, 2012; Van de Watering & Van de Rijt, 2006). These methods differ in how they split the test into halves, and whether they correct for underestimation of the reliability coefficient (Barman, 2011). Other methods, such as percent agreement, intraclass correlation coefficient, the Kappa coefficient (k), and generalisability theory (GT) analysis, can be used to determine inter-rater reliability (Downing, 2004), for example, to determine the technical correctness of ratings in the case of performance assessments (Dierick & Dochy, 2001; Downing, 2004). The statistical quality control procedure (SPC) is another approach; it is used to determine the consistency and reliability of a descriptive assessment scale (rubrics) and to identify measurement errors (Knight, Allen, & Mitchell, 2012). Determination of the standard error of measurement (SEM) is another method used to estimate reliability. The true score of a student on a test is the obtained score minus the error of measurement caused by, for example, anxiety. The error of measurement differs between students (Anderson & Rogan, 2010). To gain an indication of this error of measurement, the SEM is calculated per test, not per student (Anderson & Rogan, 2010; Burton, 2004; Hambleton & Slater, 1997).

3.4. Perspectives in the evaluation of assessment quality

This section addresses the fourth research question: Which perspectives are identified in the evaluation of assessment quality? Based on text analysis, 10 perspectives were grouped into four themes: students, staff, government, and experts (Fig. 12).

Perspectives on assessment quality differ among stakeholders because of the variations in their roles and experiences (Baartman et al., 2011; Bronkhorst, Baartman, & Stokking, 2011; Gulikers et al., 2009). Stakeholders set their own quality criteria (Zakrzewski & Steven, 2003) or use their own terminology (Sambell et al., 1997). In the evaluation of the quality of assessment programmes, the differences and similarities in the perspectives of stakeholders can provide reviewers with information about assessment quality (Dijkstra et al., 2010; Gulikers et al., 2009). Stakeholders' perceptions about assessment quality are especially important because they use the assessment in educational practice, and they experience the consequences of the assessment (Baartman et al., 2007a; Gulikers et al., 2009; Moss, 1994). Research on stakeholders' comprehension of and reactions towards an assessment may provide useful information of its meaningfulness (Linn et al., 1991). Furthermore, taking perspectives into consideration when designing assessment programmes seems useful for the acceptance of the assessment, for commitment to the assessment, for inviting creative suggestions, and for establishing fitness for practice, which is required to achieve quality (Dijkstra et al., 2010; Van der Vleuten, 1996). When constructing assessment, it is important to find a balance between different viewpoints (Frederiksen, 1989).

With regard to the perspectives of stakeholders, in literature, no insight is provided for the perspectives on the themes of assessment quality criteria, validity, transparency, and reliability. Only insights on the stakeholders' perspectives on assessment quality, in general, or on criteria within the themes of assessment quality criteria, such as authenticity or fairness, are presented.

3.4.1. Students

Students is the first theme associated with perspectives in the evaluation of assessment quality. The students' perspective is the theme most frequently mentioned (n = 130), and it is the theme that is most connected to the other perspectives. Students use the assessment, so they experience the consequences of it. It is relevant to include their perceptions on assessment in the evaluation (Baartman et al., 2007a; Kane, 2001). This section presents the perspectives of students on the theme of assessment quality criteria, transparency and the assessment quality criterion item difficulty.

With regard to authenticity, which is an aspect of theme transparency, it was found that students' learning may be unsupported, if not obstructed, when they do not perceive an assessment task as being authentic (Gulikers, Kester, Kirschner, & Bastiaens, 2008). Students respond differently to a real professional situation than to simulated situations, and when they perceive a task and the context as authentic they study more deeply (Gulikers et al., 2008). The perception of authenticity differs between students, and it may change based on their experience with professional practice (Gulikers et al., 2004; Gulikers et al., 2008). There appears to be a difference in the perceptions of freshmen and seniors (Gulikers et al., 2008), and the students' perceptions about authenticity differ from the teachers' perceptions and the developers' perceptions (Gulikers et al., 2004). The perspectives of seniors correspond more with the perspectives of teachers than those of the sophomores (Gulikers et al., 2004). According to Gulikers et al. (2004), 'Authenticity is subjective, which makes student perceptions important for authentic assessment to influence learning' (p. 69). Students also evaluate assessments on the criterion of fairness (Baartman et al., 2007a; Holmes & Smith, 2003; Sambell et al., 1997; Tillema, Leenknecht, & Segers, 2011), which is also an aspect of theme transparency. In peer assessment, students find fairness to be an important aspect of task selection (Tillema et al., 2011). Fairness is often underestimated by staff (Sambell et al., 1997), which is remarkable because fairness influences students' attitudes and judgments towards staff (Tata, 1999). Sambell et al. (1997) explained it in this way:

> To students, the concept of fairness frequently embraces more than simply the possibility (or not) of cheating: it is an extremely complex and sophisticated concept which students use to articulate their perceptions of the worth of an assessment mechanism, and it relates closely to our notions of validity. (p. 362)

From the students' perspective, the components that increase the fairness of an assessment are: relevance of the task for real (professional) life, reasonable requirements, active student involvement, development of a range of skills, appraisal of study time and effort by the assessment score, and positive long-term effects (Sambell et al., 1997). Students' perceptions on fairness differ from teachers' perceptions. In general, staff judged assessment more positively than students, and they disagree with students about the unfairness of assessment results (Meyer et al., 2010). Students want their grades to be based on facts, not opinion (Holmes & Smith, 2003).

Finally, this section will discuss the criterion of assessment quality item difficulty that does not belong to any of the overarching assessment quality themes. In Fig. 2, item difficulty is depicted by one of the grey dots outside of the three main themes. It is less frequently mentioned than other assessment quality criteria and less connected to them. According to students, an item becomes more difficult when the content is highly complex, the formulation is unclear, it contains a case study, it requires transfer to another construct domain, it is too detailed,

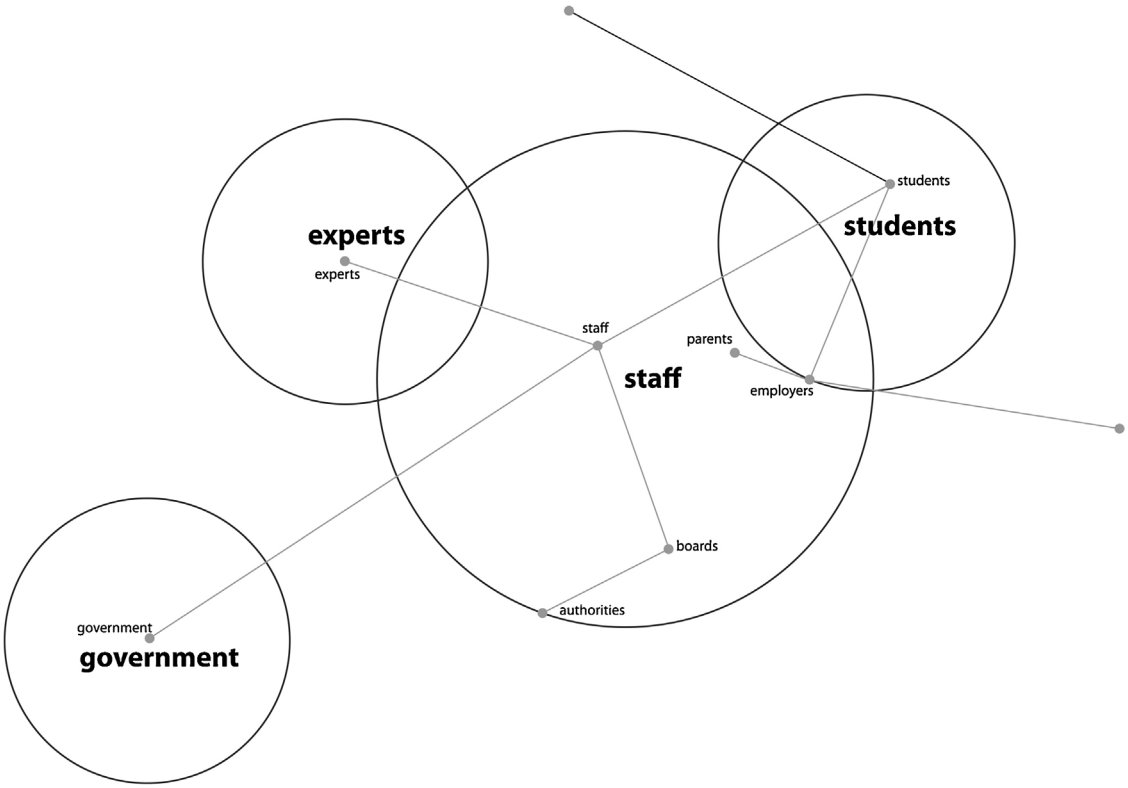

Fig. 12. The themes associated with perspectives in the evaluation of assessment quality, including the underlying perspectives. The grey dots depict the perspectives that are grouped within the themes. The lines depict their interrelations.

abstract, or long and when there appears to be more than one plausible answer (Van de Watering & Van de Rijt, 2006). Teachers underestimate the actual difficulty of items for students and students overestimate it (Van de Watering & Van de Rijt, 2006).

3.4.2. Staff

Staff of an educational organisation is the second theme associated with perspectives in the evaluation of assessment quality (n = 77). In this context, staff members include managers, teachers, examination board members, and administrative staff. They review and approve assessment quality from their own perspectives based on their roles and responsibilities (Baartman et al., 2011; Baartman, Prins et al., 2007; Zakrzewski & Steven, 2003). Since managers have to deal with the negative reactions of students and their parents, sometimes they have a tendency to be negative about the quality of assessment (Baartman, Prins et al., 2007). This section provides an overview of the perspectives of staff in the general evaluation of assessment quality.

Managers appeared to be more focused on quality assurance, and, thus, preventing mistakes, than on creating the broad outlines of a common vision for evidence-based assessment (Meyer et al., 2010). Managers and teachers expressed concerns about the restrictions that may affect assessment quality, including time, workload, and limited choices for types of assessments (Meyer et al., 2010). Instead of choosing assessments that would be most beneficial for students' learning, they chose assessments that would be more advantageous in terms of their construction, implementation, and use (Meyer et al., 2010).

Baartman et al. (2007b) found that teachers perceived older quality criteria (e.g. the reproducibility of decisions, which is an aspect of theme reliability) to be just as important as newer quality criteria (e.g. meaningfulness, which is an aspect of theme validity). They found that teachers perceive transparency to be more important than the other quality criteria, and teachers also recognise the importance of high-quality assessment programmes.

3.4.3. Government

Government is the third theme associated with perspectives in the evaluation of assessment quality (n = 14). The relevance of the government's perspective is mainly related to (legal) rules and regulations related to assessment quality, and the impact these have on other stakeholders. According to Kane (2008), 'The need for validation derives from legal . . . expectations that the claims and decisions based on test scores will be justified' (p. 79). The government's perspective can affect decisions made in educational organisations. National funding rules may lead to a situation in which schools pass students that are not on the required exit level (Meyer et al., 2010), or political developments might initiate improvements (Dijkstra et al., 2010).

3.4.4. Experts

Experts is the fourth theme associated with perspectives in the evaluation of assessment quality (n = 10). Experts are subject matter experts or experts affiliated with an external (accreditation) body (Dijkstra et al., 2010; Linn et al., 1991). Experts should be involved in constructing, reviewing, and approving assessments (Linn et al., 1991; Zakrzewski & Steven, 2003). Reviews of assessment programmes by experts outside of an organisation are generally executed to ensure accreditation and benchmarking (Dijkstra et al., 2010).

4. Discussion

By systematically summarising and organising the results of 78 peer-reviewed articles, this integrative literature review has conceptualised assessment quality in tertiary education, resulting in an overview containing:

1. The themes of assessment quality criteria: validity, transparency, and reliability;
2. The themes of influences on assessment quality criteria: standardisation, stakeholders, clarity, and construct irrelevant variance;

3. The themes associated with the evaluation of assessment quality criteria: validation and statistical data analyses;
4. The themes associated with the perspectives that should be acknowledged in the evaluation of assessment quality: students, staff, government, and experts.

The merit of this review study is that Leximancer's objective text analysis, which produces outcomes without researcher supervision (Hansson, Carey, & Kjartansson, 2010), confirms that validity, transparency, and reliability are the main themes of assessment quality criteria. This analysis also visualises the relationships among the themes, and among the concepts within each theme; thus, it provides a more complete overview of assessment quality. It is notable that, based on this analysis, all the quality criteria are clustered within the three main themes of quality criteria: validity, transparency, and reliability. This suggests that, beyond validity, transparency, and reliability, no other quality criteria need to be distinguished. These three quality criteria should be defined in terms of the purpose and proposed use of assessment, since assessment is only considered to be of quality when it is suited for its purpose and proposed use. In their study on the quality criteria of peer assessment, Ploegh et al. (2009) noted that 'peer assessment practices entail many of the quality criteria recognized in measurement and evaluation, although in an embedded way: the generic quality criteria are tuned or adapted to the setting of measurement, which is peer assessment' (p. 108). Therefore, an additional (review) study should focus on how the three assessment quality criteria, validity, transparency, and reliability, are fulfilled in tertiary educational practices. While evaluating various assessment instruments, such as rubrics or grading guides, and describing their quality, purpose and use, the differences in attaining assessment quality become visible.

Furthermore, the results provide an overview of the evolution of the concept of assessment quality in tertiary education. The results reveal that there is no uniform conceptualisation of assessment quality. In fact, there is little consensus among academics, particularly regarding to the assessment quality criteria validity and reliability. This finding is supported by the observation that more than half of the articles included in this study are opinion articles, used to provoke debate and stimulate new research. Since the present study focused on tertiary education, further research is necessary to determine whether this evolution of the concept of assessment quality is generalisable to other contexts, such as primary and secondary education. In addition, the evolution of assessment quality is affected by different kinds of economic and social-cultural developments. However, assessment quality evolved in the 20th century, and it is subject to change. Moreover, the concept of validity is still subject to ongoing debate (De la Torre, 2013). In daily practice, books, standards, and research articles, the same terminology is used; however, this does not imply that identical terms have the same meanings for different stakeholders. Thus, one implication of the present study is that stakeholders should be aware that conceptualisations might differ so they should be made explicit to avoid ambiguity.

The results show the relationships among all the influences on the assessment quality criteria. For example, standardisation in the form of assessment procedures may be influenced by how stakeholders use the procedures. One of the limitations of text analysis is that it does not reveal if any of the relations are causal or correlational, since, based on the maps provided by Leximancer, it is only possible to obtain qualitative interpretations of the data. Experimental research could be conducted to investigate the nature of the relationships among the influences and between the assessment quality criteria and the influences, as identified by the text analysis.

The results do not describe how transparency should be evaluated. Because transparency is the second most important theme of assessment quality criteria, this cannot be ignored. It is possible that this is a missing element in the existing literature; however, this finding may also be the result of using text analysis as a data reduction technique, since this excludes some of the less frequently mentioned and connected concepts. The website, www.assessmentquality.com, provides a detailed description of the concepts that are less frequently mentioned and less often connected to each other. Regardless of the reason for this result, the implication is that further research is needed to investigate how transparency can be evaluated.

Regarding the perspectives of stakeholders, there appears to be no results for perspectives on the themes of assessment quality criteria, validity, transparency, and reliability separately. Instead, the stakeholders' perspectives appear to be on assessment quality, in general, or on criteria within the themes of the assessment quality criteria, such as students' perceptions of authenticity or fairness. This is remarkable because stakeholders are most affected by inferior assessment quality. Furthermore, they are the second most important influence on assessment quality, as is shown by the results of this review study. A possible explanation for this gap is the absence of a clear conceptualisation of assessment quality, which forces researchers to either focus on specific aspects or use the concept as a whole. Another possible explanation is that stakeholders use their own terms for assessment quality. Sambell et al. (1997) showed that students use a different terminology than assessment specialists; for example, while assessment specialists use the term, validity, to refer to consequential validity students use the term, fairness. An implication of these results is that further research should focus on stakeholders' perspectives of validity, transparency, and reliability. A practical implication of the study is that educational organisations should focus on the differences in conceptualisations and perspectives among groups of stakeholders in order to develop policies and provide guidelines to optimise assessment quality in organisations.

In summary, this review study provides a clear and overarching conceptualisation of assessment quality in tertiary education. Teachers, educational advisors, and managers that are tasked with developing and implementing policies to guarantee and improve assessment quality can use the results of this study as input for conversations among students, staff, and experts. In these discussions, a consensus should be reached about what assessment quality means for the organisation and what factors should be considered in order to guarantee and achieve assessment quality. Based on those conversations and the results of this review study, teachers, educational advisors, and managers can translate their conceptualisation of assessment quality into guidelines, measures, and facilities, which can be used to assure, evaluate, monitor, and improve assessment quality in educational practice. In addition to the practical relevance, the results of this review study provide insights into the current state of scientific evidence, stimulate discussion, and offer suggestions for further research in order to improve and guarantee overall assessment quality in tertiary education.

Appendix A. Methodological Appendix

Methodological Appendix

In this methodological appendix, the methodology of the journal article 'Assessment Quality in Tertiary Education: An Integrative Literature Review' is described in detail.

Table A1
Tabulation.

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Allen et al. (2008)	Presentation of the background information for coefficient alpha and an illustration of how alpha behaves.	NE	• students (n = 104) • statistical class in the college of engineering • university • US	• statistics concepts inventory • focus group	• comparison of alpha to the average inter-item correlation • discrimatory index	+	+	−	+	+	+	−	−
Anderson and Rogan (2010)	Overview of procedures to improve the quality of assessment.	O	• ZA	–	–	+	−	−	−	+	−	+	−
Archer and McCarthy (1988)	Overview of personal biases in the assessment process.	R	• UK	–	–	+	−	−	−	+	+	+	−
Baartman et al. (2006)	Description and evaluation of a framework of quality criteria for competency assessments.	NE	• international assessment experts (n = 15) of IL, US, UK, DE, NO, NL • NL	• electronic group support system (eGSS) • expert meeting	• quantitative techniques • qualitative techniques	+	+	−	+	+	+	+	+
Baartman et al. (2007a)	Comparison of a framework consisting of 10 quality criteria for competency assessment programmes with Messick's framework of construct validity.	NE	• NL	–	–	+	−	−	−	+	+	+	+
Baartman et al. (2007b)	Exploration of the opinion of teachers on quality criteria for competency assessment programmes.	NE	• teachers (n = 211) • personal, social services, health care, economics, technology sectors • pre-vocational and vocational education • NL	• questionnaire	• Cronbach's Alpha • t-tests • ANOVA	+	+	+	+	+	+	+	+
Baartman et al. (2013)	Examination of the quality of assessment, to identify critical factors influencing assessment quality, and whether self-evaluations lead to improvement points.	NE	• teachers (n = 60), students (n = 49) • health, social, international studies, financial services, primary teacher education, construction & infrastructure, ict & business, marketing & business management, social work • higher vocational education • NL	• lecture • web-questionnaire • group interview	• Cronbach's Alpha • one sample t-tests • qualitative analysis: member checking meta-matrix	+	+	+	+	+	+	+	+
Baartman, Prins et al. (2007)	Determination of the quality of competency assessment programmes with a self-evaluation procedure.	NE	• department managers, examination board members, teachers (n = 22) • laboratory technology • vocational education • NL	• web-based evaluation tool • group interviews	• calculation of percentage of ratings • transcribing of interviews • qualitative analyses • coding	+	+	+	+	+	+	+	+
Baartman et al. (2011)	Validation of the self-evaluation procedure by comparison the outcomes of two self-evaluations.	NE	• teachers (n = 2), managers (n = 2), examination board members (n = 2) • laboratory technology • vocational education • NL	• web-questionnaire • group interviews	• cross-case comparison • verification by independent researcher	+	+	+	+	+	+	+	+
Barman (2011)	To highlight the feasibility of applying reliability analysis based on classical test theory in student assessment.	R	• Pubmed.org and Ebscohost • MY	• databases • keywords • cross references	–	+	−	−	−	+	−	−	−
Benett (1993)	Validity and reliability from the perspectives of the classical test theory, applied on assessments in the workplace.	O	• UK	–	–	+	−	−	−	+	+	+	−
Berk (1980)	Compilation and evaluation of the reliability indices in a "consumers' guide".	NE	• reliability indices (n = 13) for criterion-referenced tests • US	–	• grouping into three categories	+	−	−	+	+	+	+	−
Birenbaum (2007)	Proposition of a framework for evidence based evaluation of the quality of an assessment practice.	O	• IL	–	–	+	−	−	−	+	−	−	−
Borsboom et al. (2004)	Offering a conception of test validity.	O	• NL	–	–	+	−	−	−	+	+	+	−

(continued on next page)

Table A1 *(continued)*

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Boud (2000)	Draw attention to the double duty of assessment.	O	• AU	–	–	+	−	−	−	−	+	−	−
Bronkhorst et al. (2011)	Explanation of the quality standards in self-evaluation of assessment quality.	NE	• stakeholders • applied natural sciences and a teacher training programme • university of professional education • NL	• explanatory training • multiple choice test • questionnaire • group interview	• transcribing of group interviews • segmentation with MEPA • coding and discussion	+	+	+	+	+	+	+	+
Burton (2004)	Reliability of multiple choice and true/false tests.	O	• data on 13 negatively marked true/false tests • UK	–	–	+	−	−	+	+	−	−	−
Colliver et al. (2012)	Encourage a discussion of use of construct validity in medical education.	O	• US	–	–	+	−	−	−	+	+	+	−
Cronbach and Meehl (1955)	Identification of the concept construct validity and the elaboration of its implications.	O	• US	–	–	+	−	−	−	+	+	+	−
Dennis (2007)	Investigation of the halo effects in grading student projects.	NE	• graders (n = 32) • faculty members • psychology department • university • UK	• 502 reports were graded by two people	• preliminary analyses • regression analyses • correlation analyses	+	+	+	+	+	+	+	+
Dierick and Dochy (2001)	Operationalisation of the new assessment culture, by describing assessment forms and giving thoughts on assessment quality criteria and quality control.	NE	• NL	–	–	+	−	−	−	+	+	+	−
Dijkstra et al. (2010)	Development of design principles for assessment programmes.	NE	• experts (n = 9) programme directors, committee members • undergraduate and graduate education • US and EU • NL	• focus groups	• transcribing of focus groups • coding transcripts	+	+	+	+	+	+	+	+
Downing (2004)	Exploration of the importance of reliability in assessments.	O	• US	–	–	+	−	−	−	+	+	+	−
Downing and Haladyna (1997)	Addresses validity evidence as it relates to item development and item responses.	O	• US	–	–	+	−	−	−	+	+	+	−
Ebel (1983)	Means for the development of valid tests, and bases for defending their validity.	O	• US	–	–	+	−	−	−	+	+	+	−
Ediger (2001)	Description of the problems in grading.	O	• US	–	–	−	−	−	−	−	−	−	−
Frederiksen (1989)	Identification of characteristics that contribute to systemic validity. Development of a set of design principles for an alternative form of testing system.	O	• UK	–	–	+	−	−	−	+	−	+	+
Gulikers et al. (2004)	Exploration whether the five-dimensional framework is a complete description of authenticity and what the relative importance is of the dimensions.	NE	• students (n = 28) and teachers (n = 11) • nursing college • vocational education • NL	• literature study • session with an electronic group support system	• transcribing of discussions • qualitative analyses	+	+	+	+	+	+	+	+
Gulikers et al. (2009)	Finding empirical evidence for the theoretical characteristics of competency-based assessment and their relationship to quality criteria.	NE	• developers of the national assessment framework (n = 26) • agricultural • vocational education • NL	• questionnaires • semi-structured focus group interviews (audio taped)	• one-sample t-tests • one-way ANOVAs • games-howell post hoc corrections • method of cross-case comparison	+	+	+	+	+	+	+	+
Gulikers et al. (2008)	Examination of the perceptions of assessment authenticity and their relationship with student learning.	NE	• freshman students (n = 81) and senior students (n = 118) • social work • vocational education • NL	• four questionnaires • focus groups	• Cronbach's Alpha • confirmatory factor analysis • MANOVA • structural equation modelling • coding interrater agreement	+	+	+	+	+	+	+	+

(continued on next page)

Table A1 *(continued)*

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Haladyna et al. (2002)	Examination and evaluation of sources of evidence bearing on the validity of 31 multiple choice item-writing guidelines.	R	• textbooks (n = 27) • research studies (n = 27) 1990–2002 • item and item statistics • medical, dental, nursing, police officers, psychology and communications, biology, science • US	• two authors read passages • by disagreement a researcher reread and consensus was achieved	• validation procedure: collective judgment of two researchers about the validity of each guideline by two sources of evidence: textbook authors, empirical research	+	+	+	+	+	+	+	+
Hambleton and Murphy (1992)	Addressing the validity of several of the popular criticism of objective tests, and consideration of the viability of some of the alternatives.	O/R	• US	–	–	+	−	−	−	+	+	+	−
Hambleton and Slater (1997)	History of developments in the assessment of reliability of credentialing examinations.	NE	• performances of students (n = 1000) were simulated • candidate scores were simulated using a random number generator • US	• pass-fail decision was made • criterion scores were computed	• decision accuracy was assessed by comparing the first administration pass-fail examination decisions, and the criterion pass-fail decision	+	+	+	+	+	+	+	−
Harnisch and Mabry (1993)	Relationship between education and standardised tests in the USA and some issues in the development of alternative measures of student achievement.	O	• US	–	–	+	−	−	−	+	−	−	−
Holmes and Smith (2003)	Student perceptions of faculty grading methods.	NE	• students (n = 270) • marketing class • college of business at a midwestern university • US	• survey instruments	• review of the responses • categorisation of the responses in categories by two students • interrater reliability scores • discussion by disagreement	+	+	+	+	+	−	+	−
Kane (1992)	Description of the argument-based approach in more detail.	O	• US	–	–	+	−	−	−	+	+	+	−
Kane (2001)	Description of the history of construct validity, and summarisation of the current state of validity. Introduction of an argument-based approach.	O/R	• US	–	–	+	−	−	−	+	+	+	−
Kane (2008)	Argumentation that validation should involve an evaluation of the proposed interpretations and uses of test scores.	O	• US	–	–	+	−	−	−	+	+	+	−
Knight et al. (2012)	Development of a rubric, and demonstration of the application of quality control principles.	NE	• instructors (n = 8) • public relations instructors • midwestern college • US	• instructors judged the same papers of 7 students with Rubric A and Rubric B two times	• gage capability analysis	+	+	+	+	+	+	−	−
Knight (2000)	Explanation of a systematic approach to assessment to achieve more reliable assessments.	O	• UK	–	–	+	−	−	−	+	−	−	−
Knight (2002a)	Exploration of the use of assessment as performance indicator for quality monitoring.	O	• UK	–	–	+	−	−	−	+	+	+	−
Knight (2002b)	Insight in the importance of changing the understanding of assessment issues in higher education among stakeholders.	O	• UK	–	–	+	−	−	−	+	+	+	−
Leigh et al. (2007)	Examination of assessment models and their feasibility for professional psychology.	O/R	• US	–	–	+	−	−	−	+	−	+	−
Linn et al. (1991)	Arguments for and redefinition of quality criteria of educational (performance based) assessments.	O	• US	–	–	+	−	−	−	+	+	+	+

(continued on next page)

Table A1 *(continued)*

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Maclellan (2004)	Classification of the construct of alternative assessment through a conceptual analysis of its validity.	O	• UK	–	–	+	–	–	–	+	+	+	–
Malouff (2008)	Description of different types of biases in grading.	O	• AU	–	–	+	–	–	–	+	–	–	–
Martin (1997)	Examination of the principles of assessment and testing theory and application of the judgemental model to workplace performance.	O	• UK	–	–	+	–	–	–	–	–	–	–
Maxwell (2012)	Argument that assessment in the final year can benefit from quality assessment tasks linked to professional practice.	O	• AU	–	–	+	–	–	–	–	+	–	–
McKenna and Bull (2000)	Consideration of the general quality assurance issues of computer-assisted assessment (CAA) and of the methods which can be used to evaluate the usage of CAA.	NE	• representatives of 25 institutions • quality assurance staff • higher education • UK	–	–	+	–	–	–	+	+	–	–
Messick (1995)	Presentation of a comprehensive theory of construct validity that addresses both score meaning and social values in test interpretation and test use.	O	• US	–	–	+	–	–	–	+	+	+	–
Meyer et al. (2010)	Investigation whether and how attitudes towards, experiences with and expectations for assessment held by academic staff and their students are accommodated or represented in assessment policy and policy guidelines.	NE	• academic staff (n = 879), first year students (n = 1238), academic managers (n = 14) • tertiary institutions • NZ	• survey • qualitative research of comments and policy documents • interviews	• factor analyses • MANOVA • grounded theory approach	+	+	+	+	+	+	+	–
Moss (1994)	To illuminate and challenge the presumption that reliability, is essential to sound assessment practice.	O	• US	–	–	+	–	–	–	+	+	+	–
Moss (1995)	To highlight some of the major questions in deciding how to conceptualise validity.	O	• US	–	–	+	–	–	–	+	+	+	–
Ploegh et al. (2009)	Exploration of the quality criteria in peer assessment in classrooms, to provide a background for appraising measurement quality in 'assessment for learning'.	NE	• teachers (n = 56) • disciplines such as technology, health, economics • secondary vocational education • NL	• online questionnaire	• descriptive overview	+	+	–	+	+	+	+	–
Price et al. (2010)	Provision of a framework to examine assessment policy and practice.	O	• UK	–	–	+	–	–	–	+	+	+	–
Sambell et al. (1997)	Illustration of the impact of assessment practices on student perceptions of learning and on their learning behaviour, as an aspect of the consequential validity of assessment.	NE	• cases (n = 13) • social science, built environment, business studies, social sciences, IT/computing, IT/business, languages, design/history, psychology, professional practice studies, engineering, design • university • UK	• (group) interviewing • examination of documentary evidence	• cross-case analysis • transcription • coding	+	+	+	+	+	+	+	+
Schuwirth and Van der Vleuten (2003)	Discussion of the general issues of written assessment and overview of the used types and advantages and disadvantages.	O	• NL	–	–	+	–	–	–	+	–	–	–

(continued on next page)

Table A1 *(continued)*

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Schuwirth and Van der Vleuten (2004)	Discussion of written questions.	O	• NL	–	–	+	–	–	–	+	+	+	–
Schuwirth and Van der Vleuten (2006)	Description of the weaknesses of the current psychometric approach to assessment as a scientific model.	O	• NL	–	–	+	–	–	–	+	–	+	–
Schuwirth and Van der Vleuten (2011)	Description of the development of the idea of programmatic assessment in the context of assessment for learning.	O	• NL	–	–	+	–	–	–	+	+	+	+
Schuwirth and Van der Vleuten (2012)	Description and explanation of Kane's validity approach.	O	• AU/NL	–	–	+	–	–	–	+	–	–	–
Segers et al. (2001)	Discussion of the overall test and the problem based learning environment and descriptions of quality criteria for new modes of assessment.	NE	• students (n = 100), students (n = 48), staff members (n = 8) • economics & business administration • university • NL	• questionnaire • focus groups	• Cronbach's Alpha • data of interview was grouped in themes and structures	+	+	+	+	+	+	+	–
Segers et al. (1999)	Discussion of assessment practices within a problem based curriculum.	NE	• first year undergraduate students (n = 34) • marketing and organisation • university • NL	• knowledge test • sorting task • overall test	• Cronbach's Alpha • Pearson correlation coefficient	+	+	+	+	+	+	+	+
Shepard (1993)	Description of the evolution of test validity.	O/R	• US	–	–	+	–	–	–	+	+	+	–
Spence-Brown (2001)	Exploration of the theoretical and practical issues surrounding authenticity in assessment.	NE	• cases (n = 8) • intermediate level • Japanese course • university • AU	• case study method • assessment items • marks and comments of teachers • retrospective interviews	–	+	+	–	+	+	+	+	–
Stowell (2004)	Argument of aspects of the equality and standards agendas in the UK Higher Education.	O	• UK	–	–	+	–	–	–	–	+	+	–
Tata (1999)	Examination of the connections between students' evaluation of instructors, the fairness of grade distributions, the fairness of grading procedures, and evaluations of the instructor.	E	• undergraduate students (n = 97) • management • university • US	• participants randomly assigned to one of the 4 manipulation conditions. • reading of scenario • questionnaire	• Cronbach's Alpha • t-tests • ANOVA • partial correlation coefficients	+	+	+	+	+	+	+	+
Tillema et al. (2011)	Identification of quality criteria usable for the design of (peer) assessment for learning.	R	• journal articles (n = 42) • studies published 1990–2007 • Econlit, ERIC, PsychINFO, Web of Science • NL	• two searches in databases by keywords • selection criteria • inclusion criteria	• mind maps • narrative summaries	+	+	+	+	+	+	+	–
Tweed and Wilkinson (2012)	Application of the principles of evaluating diagnostic tests to educational assessment.	O	• NZ	–	–	+	–	–	–	+	–	+	+

(continued on next page)

Table A1 *(continued)*

Study	Focus of the research	Design	Participants characteristics | literature characteristics | country	Data collection	Data analysis	Critical appraisal							
						a	b	c	d	e	f	g	h
Van de Watering and Van de Rijt (2006)	Item difficulty in higher education from the perspective of constructors, assessment composers, and students.	NE/R	• ERIC, ISI Web of Knowledge, Science Direct, Online Contents, Google Scholar • teachers (n = 17), first year students (n = 223), first year students (n = 138), first year students (n = 198), first year students (n = 30) • faculty of law • university • NL	• search in databases by keywords • snowball method • three assessments with 41 questions • construction meetings • two questionnaires • focus group	• data analyses literature search and focus groups not described • Cronbach's Alpha • p-values of assessment items • rating scale analyses	+	+	−	+	+	+	+	−
Van der Vleuten (1996)	Development, research, and practical implications of the assessment of professional competence.	O	• NL	–	–	+	−	−	−	+	+	+	−
Van der Vleuten et al. (1991)	Examination of the claim that objectified measures produce better reliability.	O/R	• NL	–	–	+	−	−	−	+	+	+	−
Van der Vleuten et al. (2012)	Presentation of a model for programmatic assessment.	O	• NL	–	–	+	−	−	−	+	+	+	−
Van der Vleuten and Schuwirth (2005)	Argument that a conceptual model for defining the utility of an assessment, can serve as a guide to the design of assessment programmes.	O	• NL	–	–	+	−	−	−	+	+	+	−
Verhoeven et al. (1999)	Illustration of the workings of test committees and the process used to produce progress tests.	O	• NL	–	–	+	−	−	−	+	+	+	+
Woolf (2004)	Discussion how teachers might optimise the use of assessment criteria.	NE	• institutions (n = 7) • undergraduate history and business programmes • student assessment and classification working group (SACWG) • university • UK	• case study of the assessment and grading criteria within project handbooks	• comparison of assessment criteria and grading criteria	−	+	−	−	−	−	+	+
Wools et al. (2010)	Illustration of a procedure for the evaluation of validity and validation based on the argument-based approach.	O	• NL	–	–	+	−	−	−	+	+	+	−
Zakrzewski and Steven (2003)	Proposition that The Catherine wheel' risk model for Computer-based assessment (CBA) has a synergy with stakeholders' view on quality and is of use for a successful implementation of CBA.	O	• UK	–	–	+	−	−	−	−	+	+	−

Note. E = experimental research | NE = non-experimental research | O = opinion | R = review | AU = Australia | DE = Germany | EU = European Union | IL = Israel | MY = Malaysia | NL = The Netherlands | NO = Norway | NZ = New Zealand | UK = United Kingdom | US = United States of America | ZA = South Africa.

[a] The problem formulation gives a clear statement of the objectives and the scope of the study (American Educational Research Association, 2006; Spencer, Ritchie, Lewis, & Dillon, 2003).

[b] The process of data collection is precisely and transparent described and argued (American Educational Research Association, 2006; Spencer et al., 2003).

[c] The process of data analyses is precisely and transparent described and argued (American Educational Research Association, 2006; Spencer et al., 2003).

[d] Results are supported by the required research evidence (American Educational Research Association, 2006; Spencer et al., 2003).

[e] Results are clearly linked to the research goals (Spencer et al., 2003).

[f] The conclusion is clearly linked to the research goals/question(s) (American Educational Research Association, 2006; Spencer et al., 2003).

[g] The theoretical, practical, or methodological implications are discussed (American Educational Research Association, 2006; Spencer et al., 2003).

[h] Limitations of the research in meeting the goal(s), in the design, or in acquired evidence are discussed (Spencer et al., 2003).

Search method

The research database EBSCO (Academic Search Elite, Business Source Premier, E-Journals, GreenFILE, Library Information Science & Technology Abstracts, PsychINFO, Regional Business News, ERIC, Psychology, and Behavioral Sciences Collection, PsycArticles) was systematically searched. The search path presented in Fig. 1 shows the search terms that were used related to assessment and quality, and the specifications of the context. Only articles focusing on tertiary education, written in English, and published in peer-reviewed journals from 1998 to 2014, were used. This resulted in 396 hits after removal of duplicates.

To be included in this review, a study had to address and operationalise assessment quality in tertiary education, so all the abstracts were read to verify this. A total of 33 articles remained. Using the snowball method (Fig. 1), in which the reference list of each publication was screened for additional articles related to assessment quality, a total of 91 articles were found. The abstracts of these 91 articles were read to determine whether the articles focused on tertiary education and operationalised assessment quality, and were published in peer-reviewed journals. Following the application of these selection criteria, a total of 45 articles remained, of which 35 fully complied with the inclusion criteria. Following further analysis of the content, 10 articles remained. These peer-reviewed articles operationalised assessment quality, in general, but they used examples of all types of education, including tertiary education (e.g. Kane, 2001), referred to test forms that are applicable in tertiary education (e.g. Messick, 1995), or operationalised one of the key concepts of assessment quality (e.g. Borsboom et al., 2004). One consequence of including these articles might be that the review study lost some of its intended focus on assessment quality in tertiary education. One advantage of including these articles is that they provide a wider overview of the development of assessment quality over time, since these articles represent the foundation of the current body of publications of assessment quality in tertiary education.

A total of 78 journal articles were further analysed. Of these articles, 41 were opinion articles, 26 were non-experimental articles, 1 was an experimental article, 5 were opinion/review articles, 1 was a non-experimental/review article, and 4 were review articles. The articles originated from Australia, Israel, Malaysia, New Zealand, South Africa, the Netherlands, the United Kingdom, and the United States. Table A1 presents information about each of the journal articles included in this review study.

Synthesis of the studies

Since the focus of this integrative literature review is an overview of research on assessment quality, the qualitative framework synthesis approach was used (Carroll et al., 2011; Dixon-Woods, 2011; Gough et al., 2012). The information in each journal article was systematically and explicitly aggregated in a framework based on the research questions (Dixon-Woods, 2011). The framework consisted of three parts. Part one, *descriptive information of the journal article,* was used to systematically summarise each article using the same criteria and to provide a schematic overview to increase the transparency of the review. This is called tabulation as presented in Table A1 (Petticrew & Roberts, 2006). The practical and theoretical relevance of and foundations for each journal article were summarised and analysed to interpret the article within its own context (Greenhalgh et al., 2005). Part two, *relevant information of the journal article,* was used to systematically interpret, combine, and identify information relevant to answering the four research questions. Part three, *critical appraisal*, was used to analyse the methodological quality of the included articles (Whittemore & Knafl, 2005), and to detect methodological issues related to the research topic in order to evaluate the overall quality of the review (Cooper, 1998). The critical appraisal criteria were derived from a qualitative research appraisal framework (Spencer et al., 2003) as well as standards for research (AERA, 2006). Since this is an integrative review that sought to provide an overview of the topic under study, the quality of the articles is described and discussed in Table A1, and none of the articles were initially excluded based on quality. A minimal level of quality for the articles was assured by selecting only peer-reviewed articles. Moreover, exclusion based on the quality of journal articles is difficult since there appears to be no consensus about the inclusion or exclusion criteria (Cooper, 1998; Dixon-Woods, 2006). The authors established the current framework by testing two earlier versions of the framework; they coded five journal articles, discussed the completeness, comprehensibility and accuracy of the codes and the comparability of the coded segments, and then made improvements.

The first author coded all of the 78 journal articles using the coding framework and the data-analysis software MAXQDA version 11. The second author evaluated the coded segments in part 2 of the coding framework. The comprehensibility of the 2117 coded segments was verified as a precondition for judgment of the content. Then, the selected segments were judged on content. The intention of that process was to assess the quality of the categorisation of the segments by approving the relevance and accuracy of the coding, or to challenge it by offering an alternative code (Carroll et al., 2011). Another considered alternative was double blind coding; however, this was not feasible due to the number of articles included in the review study. The second author assessed 99.6% of all segments as comprehensible, 99.8% as relevant for answering the research question and 98.8% as being coded accurately.

The authors used Leximancer 3 to execute a conceptual analysis of all the selected text segments, to determine the presence and frequency of concepts in the text and, as a relational analysis, to determine how the identified concepts were related to one another (Leximancer, 2008). Leximancer determines whether concepts are connected via proximity analysis, which determines the co-occurrence of concepts within a text. This makes the discovered relationships visible through cognitive mapping (Leximancer, 2008). Normal and default pre-processing options were used. Automatic text analysis was explored but not used, as this resulted in words not related to the code. For example, one automatically selected concept for research question 1 (RQ1) was criterion, which is not a criterion of assessment quality. Thus, the first two researchers manually determined and imported the list of concepts and the initial thesaurus terms. The following procedure was used to determine the concepts and the initial thesaurus terms. First, MAXQDA provided the word frequencies of the codes related to, respectively, RQ1, RQ2, RQ3, and RQ4. Second, the first and second authors independently selected the words related to the code. For example, transparency is a word related to code RQ1, assessment quality criteria. The percentage of agreement on codes was 88% for RQ1, 80% for RQ2, 86% for RQ3 and 95% for RQ4. If the two authors disagreed about concepts, they judged them on relevance by reviewing the word in the context of the original text segments. Third, the first author classified the words as concepts and/or as initial thesaurus terms. For example, for RQ1 the concept, validity, was selected with these initial thesaurus terms: validities, invalid, and validity. Fourth, the second author verified the classifications of the first author, and differences were discussed until consensus was reached.

Leximancer highlighted the relationships in the data and clustered the concepts into themes. Leximancer provided maps based on the text segments per code. This is depicted in Figs. 2 and 3. As seen in Fig. 2, the assessment quality criteria (grey dots) are clustered into three themes. The size of the themes was set to 50%, which means that 50% of all the criteria that were most frequently connected were clustered into one theme. The other criteria were distributed over the other themes. Fig. 3 depicts the assessment quality criteria within the theme validity and their

interrelationships. A comparable process was followed for RQ2, RQ3, and RQ4. Since text analysis was used as a data reduction technique, not all the criteria, influences, evaluations, and perspectives will be discussed in the paper to enhance the comprehensibility and readability. The figures are depicted in detail on the website: www.assessmentquality.com. After data reduction with Leximancer, five of the selected articles did not provide content that was relevant to the research questions (Dennis, 2007; Ediger, 2001; Malouff, 2008; Maxwell, 2012; Segers et al., 1999).

References[1]

Association of Educational Assessment (2012). *The 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, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Allen*, K., Reed-Rhoads, T., Terry, R. A., Murphy, T. J., & Stone, A. D. (2008). Coefficient alpha: An engineer's interpretation of test reliability. *Journal of Engineering Education, 97*, 87–94. http://dx.doi.org/10.1002/j.2168-9830.2008.tb00956.x.

American Educational Research Association (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. http://dx.doi.org/10.3102/0013189x035006033.

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. http://dx.doi.org/10.1002/bmb.20362.

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

Assessment reform group (2002). *Assessment for learning: 10 principles.* Retrieved from www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf.

Attard, A., Di Ioio, E., Geven, K., & Santa, S. (2010). *Student centered learning. An insight into theory and practice.* Retrieved from https://www.esu-online.org/wp-content/uploads/2016/07/2010-T4SCL-Stakeholders-Forum-Leuven-An-Insight-Into-Theory-And-Practice.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. http://dx.doi.org/10.1016/j.stueduc.2006.04.006.

Baartman*, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2011). Self-evaluation of assessment programs: A cross-case analysis. *Evaluation and Program Planning, 34*, 206–216. http://dx.doi.org/10.1016/j.evalprogplan.2011.03.001.

Baartman*, L. K. J., Gulikers, J., & Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education, 38*, 978–997. http://dx.doi.org/10.1080/02602938.2013.771133.

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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.tate.2006.04.043.

Baartman*, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determing the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258–281. http://dx.doi.org/10.1016/j.stueduc.2007.07.004.

Barman*, A. (2011). Feasibility of applying classical test theory in testing reliability of student assessment. *International Medical Journal, 18*, 110–113 Retrieved from http://connection.ebscohost.com/c/articles/63045760/feasibility-applying-classical-test-theory-testing-reliability-student-assessment.

Benett*, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education, 18*, 83–94. http://dx.doi.org/10.1080/0260293930180201.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25. http://dx.doi.org/10.1080/0969594x.2010.513678.

Berk*, R. A. (1980). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17*, 323–349. http://dx.doi.org/10.1111/j.1745-3984.1980.tb00835.x.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364. http://dx.doi.org/10.1007/BF00138871.

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. http://dx.doi.org/10.1016/j.edurev.2006.01.001.

Birenbaum*, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*, 29–49. http://dx.doi.org/10.1016/j.stueduc.2007.01.004.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7–74. http://dx.doi.org/10.1080/0969595980050102.

Black, P., & William, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*, 623–637. http://dx.doi.org/10.1080/0141192032000133721.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of pupil learning.* New York, NY: McGraw-Hill.

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*, 655–670. http://dx.doi.org/10.1080/03075071003777716.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*, 466–481. http://dx.doi.org/10.1080/02602938.2015.1024607.

Borsboom*, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. http://dx.doi.org/10.1037/0033-295X.111.4.1061.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*, 399–413. http://dx.doi.org/10.1080/02602930600679050.

Boud, D., & Associates (2010). *Assessment 2020: Seven propositions for assessment reform in higher education.* Retrieved from https://www.uts.edu.au/sites/default/files/Assessment-2020_propositions_final.pdf.

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

Bridges, P., Bourdillon, B., Collymore, D., Cooper, A., Fox, W., Haines, C., ... Yorke, M. (1999). Discipline-related marking behaviour using percentages: A potential cause of inequity in assessment. *Assessment & Evaluation in Higher Education, 24*, 285–300. http://dx.doi.org/10.1080/0260293990240303.

Bronkhorst*, L. H., Baartman, L. K. J., & Stokking, K. M. (2011). The explication of quality standards in self-evaluation. *Assessment in Education: Principles, Policy & Practice, 19*, 357–378. http://dx.doi.org/10.1080/0969594x.2011.570731.

Burton*, R. F. (2004). Multiple choice and true/false tests: Reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education, 29*, 585–595. http://dx.doi.org/10.1080/0260293041000168915 3.

Bybee, R. W. (1997, October). The sputnik era: Why is this educational reform different from all other reforms? Symposium conducted at the meeting of Center for Science, Mathematics, and Engineering Education, Washington, DC.

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. http://dx.doi.org/10.1186/1471-2288-11-29.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16*(8), 16–20. http://dx.doi.org/10.3102/0013189x016008016.

Colliver*, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity … And back? *Medical Education, 46*, 366–371. http://dx.doi.org/10.1111/j.1365-2923.2011.04194.x.

Cooper, H. (1998). The data evaluation stage. In C. D. Laughton (Ed.). *Synthesizing research* (pp. 78–103). London, England: Sage publications.

Cronbach*, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. http://dx.doi.org/10.1037/h0040957.

De la Torre, J. (Ed.) (2013). Validity [Special issue]. *Journal of Educational Measurement, 50*(1).

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

Dierick*, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307–329. http://dx.doi.org/10.1016/S0191-491X(01)00032-3.

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. http://dx.doi.org/10.1007/s10459-009-9205-z.

Dixon-Woods, M. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research, 6*, 27–44. http://dx.doi.org/10.1177/1468794106058867.

Dixon-Woods, M. (2011). Using framework-based synthesis for conducting reviews of qualitative studies. *BMC Medicine, 9*, 39–40. http://dx.doi.org/10.1186/1741-7015-9-39.

Downing*, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61–82. http://dx.doi.org/10.1207/s15324818ame1001_4.

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

Driscoll, M. P. (2005). *Psychology of learning for instruction.* Boston, MA: Pearson Education.

Ebel*, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7–10. http://dx.doi.org/10.1111/j.1745-3992.1983.tb00688.x.

Ediger*, M. (2001). Problems in grading based on testing university students. *College Student Journal, 36*(1), 37. Retrieved from http://eric.ed.gov/?id=ED452262.

Frederiksen*, J. R. (1989). A systems approach to educational testing. *Educational*

[1] References marked with an asterisk indicate studies included in the integrative literature review.

Researcher, 18(9), 27–32. http://dx.doi.org/10.3102/0013189x018009027.
Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 5*(1), 3–31. http://dx.doi.org/10.1007/978-3-8348-9837-1.
Gibbs, G. (2010). Does assessment in open learning support students? *Open Learning: The Journal of Open, Distance and E-learning, 25*, 163–166. http://dx.doi.org/10.1080/02680511003787495.
Gilbert, F., & Maguire, G. (2014). *Assignment brief design guidelines.* Retrieved from http://assignmentbriefdesign.weebly.com/.
Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 28*(1), 1–9. http://dx.doi.org/10.1186/2046-4053-1-28.
Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science & Medicine, 61*, 417–430. http://dx.doi.org/10.1016/j.socscimed.2004.12.001.
Gulikers*, J., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*, 67–86. http://dx.doi.org/10.1007/BF02504676.
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. http://dx.doi.org/10.1016/j.learninstruc.2007.02.012.
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. http://dx.doi.org/10.1016/j.stueduc.2009.05.002.
Haladyna*, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. http://dx.doi.org/10.1207/s15324818ame1503_5.
Hambleton*, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*, 1–16. http://dx.doi.org/10.1207/s15324818ame0501_1.
Hambleton*, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19–28. http://dx.doi.org/10.1207/s15324818ame1001_2.
Hansson, T., Carey, G., & Kjartansson, R. (2010). A multiple software approach to understanding values. *Journal of Beliefs & Values, 31*, 283–298. http://dx.doi.org/10.1080/13617672.2010.521005.
Harnisch*, D. L., & Mabry, L. (1993). Issues in the development and evaluation of alternative assessments. *Journal of Curriculum Studies, 25*, 179–187. http://dx.doi.org/10.1080/0022027930250207.
Harvey, L., & Green, D. (1993). Defining quality. *Assessment & Evaluation in Higher Education, 18*, 9–34. http://dx.doi.org/10.1080/0260293930180102.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. http://dx.doi.org/10.3102/003465430298487.
Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* Abingdon, England: Routledge.
Holmes*, L. E., & Smith, L. J. (2003). Student evaluations of faculty grading methods. *Journal of Education for Business, 78*, 318–323. http://dx.doi.org/10.1080/08832320309598620.
Kane*, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. http://dx.doi.org/10.1037/0033-2909.112.3.527.
Kane*, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x.
Kane*, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*, 76–82. http://dx.doi.org/10.3102/0013189x08315390.
Knight*, J., Allen, S., & Mitchell, A. M. (2012). Establishing consistency measurements of grading for multiple section courses. *Journal of the Academy of Business Education, 13*(1), 28–47 Retrieved from http://www.abeweb.org/#!journal/c1iwz.
Knight*, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment & Evaluation in Higher Education, 25*, 237–251. http://dx.doi.org/10.1080/713611434.
Knight, P. T. (2001). A briefing on key concepts formative and summative, criterion & norm-referenced assessment. In B. R. Smith Blackwell, & M. Yorke (Eds.). *Assessment Series No. 7* York, England: Learning and Teaching Support Network Retrieved from http://neillthew.typepad.com/files/id7_briefing_on_key_concepts.rtf.
Knight*, P. T. (2002a). The achilles' heel of quality: The assessment of student learning. *Quality in Higher Education, 8*, 107–115. http://dx.doi.org/10.1080/1353832020127506.
Knight*, P. T. (2002b). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*, 275–286. http://dx.doi.org/10.1080/03075070220000662.
Leigh*, I. W., Smith, I. L., Bebeau, M. J., Lichtenberg, J. W., Nelson, P. D., Portnoy, S., ... Kaslow, N. J. (2007). Competency assessment models. *Professional Psychology: Research and Practice, 38*, 463–473. http://dx.doi.org/10.1037/0735-7028.38.5.463.
Leximancer (2008). *Leximancer the why, not just the what. Leximancer manual version 2.0.* Brisbane, Australia: Author.
Linn*, R. L., Bakker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21. http://dx.doi.org/10.3102/0013189x020008015.
Maclellan*, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education, 29*, 311–321. http://dx.doi.org/10.1080/0260293042000188267.
Malouff*, J. (2008). Bias in grading. *College Teaching, 56*, 191–192. http://dx.doi.org/10.3200/ctch.56.3.191-192.
Martin*, S. (1997). Two models of educational assessment: A response from initial teacher education: If the cap fits… *Assessment & Evaluation in Higher Education, 22*, 337–343. http://dx.doi.org/10.1080/0260293970220307.
Martinez, M. E., & Lipson, J. I. (1989). Assessment for learning. *Educational Leadership, 46*(7), 73–75 Retrieved from http://www.ascd.org/publications/educational-leadership/apr89/vol46/num07/toc.aspx.
Maxwell*, T. W. (2012). Assessment in higher education in the professions: Action research as an authentic assessment task. *Teaching in Higher Education, 17*, 686–696. http://dx.doi.org/10.1080/13562517.2012.725220.
McKenna*, C., & Bull, J. (2000). Quality assurance of computer-assisted assessment: Practical and strategic issues. *Quality Assurance in Education, 8*, 24–32. http://dx.doi.org/10.1108/09684880010312659.
Messick*, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741.
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. http://dx.doi.org/10.1111/j.1468-2273.2010.00459.x.
Ministry of Science Technology, & Innovation (2005). *A framework for qualifications of the european higher education area.* Retrieved from http://www.ond.vlaanderen.be/hogeronderwijs/bologna/documents/050218_QF_EHEA.pdf.
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. http://dx.doi.org/10.1207/s15366359mea0302_1.
Moss*, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12. http://dx.doi.org/10.3102/0013189x023002005.
Moss*, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*(2), 5–13. http://dx.doi.org/10.1111/j.1745-3992.1995.tb00854.x.
Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences.* Oxford, England: Blackwell Publishing.
Ploegh*, K., Tillema, H. H., & Segers, M. S. R. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*, 102–109. http://dx.doi.org/10.1016/j.stueduc.2009.05.001.
Price*, M., Carroll, J., O'Donovan, B., & Rust, C. (2010). If I was going there I wouldn't start from here: A critical commentary on current assessment practice. *Assessment & Evaluation in Higher Education, 36*, 479–492. http://dx.doi.org/10.1080/02602930903512883.
QAA (2014). *Assessment of students and the recognition of prior learning. The UK Quality Code for Higher Education Glaucester, England: The Quality Assurance Agency for Higher Education.* Retrieved from http://www.qaa.ac.uk/assuring-standards-andquality/the-quality-code.
Reynolds, C. R., Livingston, R. B., & Willson, V. (2010). Reliability for teachers. In L. Reinkober (Ed.). *Measurement and assessment in education* (pp. 90–122). Upper Saddle River, NJ: Pearson Education.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144. http://dx.doi.org/10.1007/bf00117714.
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. http://dx.doi.org/10.1016/S0191-491X(97)86215-3.
Schuwirth*, L. W. T., & Van der Vleuten, C. P. M. (2003). Abc of learning and teaching in medicine, written assessment. *British Medical Journal, 326*, 643–645. http://dx.doi.org/10.1136/bmj.326.7390.643.
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. http://dx.doi.org/10.1111/j.1365-2929.2004.01916.x.
Schuwirth*, L. W. T., & Van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*, 296–300. http://dx.doi.org/10.1111/j.1365-2929.2006.02405.x.
Schuwirth*, L. W. T., & Van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33*, 478–485. http://dx.doi.org/10.3109/0142159x.2011.565828.
Schuwirth*, L. W. T., & Van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education, 46*, 38–48. http://dx.doi.org/10.1111/j.1365-2923.2011.04098.x.
Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.). *Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally.
Segers*, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students knowledge profiles in a problem-based curriculum. *Learning Environments Research, 2*, 191–213. http://dx.doi.org/10.1023/a:1009932125947.
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 Education, 16*, 569–588. http://dx.doi.org/10.1007/BF03173198.
Shepard*, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450. http://dx.doi.org/10.3102/0091732x019001405.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. http://dx.doi.org/10.3102/0013189x029007004.
Spence-Brown*, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*, 463–481. http://dx.doi.org/10.1177/026553220101800408.
Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence.* Retrieved from http://www.alnap.org/resource/10033.
Stiggins, R. (2007). Assessment through the student's eyes. *Educational Leadership, 64*(8), 22–26 Retrieved from http://www.ascd.org/publications/educational-leadership/

may07/vol64/num08/Assessment-Through-the-Student's-Eyes.aspx.
Stobart, G. (2008). *Testing times. The uses and abuses of assessment.* Oxon, England: Routledge.
Stowell*, M. (2004). Equity, justice and standards: Assessment decision making in higher education. *Assessment & Evaluation in Higher Education, 29*, 495–510. http://dx.doi.org/10.1080/0260293031000168905.
Tata*, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: A justice perspective. *Journal of Psychology, 133*, 263–271. http://dx.doi.org/10.1080/00223989909599739.
Tillema*, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning – a review of research studies. *Studies in Educational Evaluation, 37*, 25–34. http://dx.doi.org/10.1016/j.stueduc.2011.03.004.
Tweed*, M., & Wilkinson, T. (2012). Diagnostic testing and educational assessment. *The Clinical Teacher, 9*, 299–303. http://dx.doi.org/10.1111/j.1743-498X.2012.00567.x.
Van Merriënboer, J. J. G., & Kirschner, P. A. (2007). A new approach to instruction. In L. Akers (Ed.). *Ten steps to complex learning* (pp. 3–11). London, England: Lawrence Erlbaum associates publishers.
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. http://dx.doi.org/10.1016/j.edurev.2006.05.001.
Van der Vleuten*, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317. http://dx.doi.org/10.1111/j.1365-2929.2005.02094.x.
Van der Vleuten*, C. P. M., Norman, G. R., & Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25*(2), 110–118. http://dx.doi.org/10.1111/j.1365-2923.1991.tb00036.x.
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. http://dx.doi.org/10.3109/0142159X.2012.652239.
Van der Vleuten*, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*, 41–67. http://dx.doi.org/10.1007/BF00596229.
Verhoeven*, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (1999). Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health: Change in Learning & Practice, 12*(1), 49–60. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=afh&AN=2012333&site=ehost-live.
Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Science, 33*, 283–309. http://dx.doi.org/10.1007/s11251-005-3002-8.
Whittemore, R., & Knafl, K. (2005). Methodological issues in nursing research. The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*, 546–553. http://dx.doi.org/10.1111/j.1365-2648.2005.03621.x.
Woolf*, H. (2004). Assessment criteria: Reflections on current practices. *Assessment & Evaluation in Higher Education, 29*, 479–493. http://dx.doi.org/10.1080/0260293031000168904.
Wools*, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 1*, 63–82. http://dx.doi.org/10.3280/CAD2010-001007.
Yang, L., & McCall, B. (2014). World education finance policies and higher education access: A statistical analysis of world development indicators for 86 countries. *International Journal of Educational Development, 35*, 25–36. http://dx.doi.org/10.1016/j.ijedudev.2012.11.002.
Yorke, M. (2008). *Grading student achievement in higher education: Signals and shortcomings.* London, England: Routledge.
Zakrzewski*, S., & Steven, C. (2003). Computer-based assessment: Quality assurance issues, the hub of the wheel. *Assessment & Evaluation in Higher Education, 28*, 609–623. http://dx.doi.org/10.1080/0260293032000130243.

Karin J. Gerritsen-van Leeuwenkamp is a PhD-student at the Welten institute of the Open University of the Netherlands and educational advisor at Saxion, University of Applied Sciences.

Desirée Joosten-ten Brinke is an associate professor in testing and assessment at the Welten institute of the Open University of the Netherlands and an associate professor in testing and assessment at Fontys, University of Applied Sciences for teacher trainers.

Liesbeth Kester is full professor Educational Sciences at the Department of Education & Pedagogy, Utrecht University.