Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning – A review of research studies

Harm Tillema a,*, Martijn Leenknecht a, Mien Segers b

a Leiden University, The Netherlands
b Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Available online 14 April 2011

Keywords:
Peer assessment
Quality criteria
Student involvement
Assessment cycle
Assessment for learning

ABSTRACT

The interest in assessment for learning (AfL) has resulted in a search for new modes of assessment that are better aligned to students’ learning how to learn. However, with the introduction of new assessment tools, also questions arose with respect to the quality of its measurement. On the one hand, the appropriateness of traditional, psychometric criteria is being questioned in the context of assessment for learning. On the other hand, it has been argued that new criteria need to be formulated to acknowledge the unique features of assessment for learning. The issue dealt with in this review is what quality criteria are specifically relevant to AfL. Studies using (peer) assessment for learning were evaluated with regard to use of quality criteria under two perspectives: their recognition of educational measurement criteria, as well as their consideration of student involvement in the assessment of learning. The selected studies were analyzed to determine what specific criteria were utilized in (successive parts of) the assessment cycle (i.e., the process of construction, administration and follow up of an assessment). Our results indicate that quality criteria are dissimilarly connected to the distinct steps of an assessment cycle, which holds as well for the attention given to student involvement in assessment for learning.

© 2011 Elsevier Ltd. All rights reserved.

Assessing the assessment

Assessing students’ learning how to learn is viewed as an essential ingredient in effective and motivating instruction (Shepard, 2000); but what are relevant criteria to determine the quality of such an assessment? Assessment for learning (Assessment Reform Group, 1999; Black & William, 1998), interpreted as providing (in)formative feedback (Popham, 2004), is regarded as a key route in accomplishing significant improvements in students' ability in learning how to learn. To promote such learning, assessment’s prime function is to endorse adaptive, student focused feedback on the learning progress of the learner (Birenbaum, 2007; Dochy & McDowell, 1997). In consequence, new modes of assessment conducive to such a promotion of learning have been called for (CCSSO, 2009). These new modes of assessment for learning are meant to scaffold coherent, authentic, personalized, direct, and practical (i.e., feasible) information to the learners (McMillan, 2007) that will help them to actively engage in successful learning activities (Biggs, 1999; James et al., 2007). As tools of assessment, they are regarded primarily to facilitate reaching improved learning outcomes in contrast to being ‘just’ measurement tools for student achievement in a summative sense (Birenbaum, 1996; Dierick & Dochy, 2001; Sadler, 2010). Various authors have stressed the importance of firmly embedding or integrating these assessment tools in the teaching and learning process (Segers, Dochy, & Cascallar, 2004; Shepard, 2000).

Quality of assessments for learning

Together with this changing perspective on assessment purposes (i.e., from assessment of learning to assessment for learning), a debate started (Birenbaum, 2007; Martin, 1997; Stiggins, 2002) on the need for relevant criteria to establish the quality of measurement in new assessment modes. It has been assumed that the use of traditional criteria of validity and reliability are no longer feasible or even relevant to appraise formative assessment tools aimed for learning (Linn, Baker, & Dunbar, 1991; Martin, 1997; Messick, 1994). A possible conflict was noted (Boud & Falchikov, 2006; Shepard, 2000; Stiggins, 2002; Yorke, 2003) between assessment tools intending to adhere to both aims and quality standards that hold on to formative promotion of learning. In the meantime, several proposals for quality criteria to support and warrant assessment construction have been generated to account for the unique features and goals of assessment for learning (e.g. Baartman, Prins, Kirschner, & Van der Vleuten, 2007; Dierick & Dochy, 2001; Frederiks & Collins, 1989). But despite the ongoing debate on assessment quality to date (Dochy, 2001), it is still unclear which specific quality criteria are to be associated
with, or taken into account when using, assessment tools for learning (Ploegh, Tillema, & Segers, 2009). Specification of the quality criteria used in the specific context of assessment for learning might help to evaluate what is considered relevant and viewed upon as applicable to scaffold student learning how to learn (Chan & Elliott, 2004). For this reason, our goal is to gauge quality criteria as they have been used in actual studies on assessment for learning to determine the adherence to and warranty of criteria in the design of assessment tasks.

In this respect, peer assessment, as one of the modes of assessment particularly conducive to a learner-centered approach on assessment, is of relevance to our purpose since it actively involves students in the assessment of their learning how to learn (James et al., 2007). Peer assessment is defined (Falchikov, 2005) as the process whereby students are involved in grading the work of their own peers (Topping, 2009). It is often stressed that the main goal of peer assessment is not to offer a final mark or evaluation, but also whether it allows for student involvement feedback and appraisal (Boud, 1995). It also is found to be an effective way to deepen understanding of students' own learning (Sivan, 2000).

In our analytical review, we aim to identify quality assurance criteria in use with regard to the design of assessment for learning tasks. For our purpose, this implies two considerations: On the one hand, it means looking into the appropriateness of how “traditional”, i.e., psychometric quality criteria are being met within the context of assessment for learning. On the other hand, we look into how assessments are constructed to enhance learning, i.e., by appraising the way students are involved in the process of assessment (Sadler, 2010). Our search then is twofold: not only needs to be established whether assessments for learning comply with standards of measurement, but also whether it allows for student involvement or participation in assessment for learning (Natriello, 1987). We, therefore, gauged whether both ends were met in studies on (peer) assessment by constructing an analytical framework that relates measurement quality criteria as well as student involvement relative or sensitive to the different phases or stages in the construction and administration of an assessment task/assignment. In this way, an overview can be provided to detect the degree to which studies, which have implemented peer assessment for learning, adhere to, or are in accordance with, quality criteria considerations as can be found in the educational measurement literature (Brennan, 2006); as well as acknowledge the students’ involvement in the assessments. This is specified with regard to the different phases of an assessment cycle. In order to develop such a framework we first identified the building blocks in the construction and delivery, i.e., design of an assessment task to connect them with quality criteria as identified in the measurement literature.

### Assessing the assessment: construction of an analytical framework

#### Assessment cycle

From literature on test construction and assessment (e.g. Brualdi, 1998; Freeman & Lewis, 1998; Messick, 1994; Moskal, 2003) we extracted seven building blocks that usually are identified as covering the different steps in the construction, delivery and decision making of assessment tasks. This process is also referred to as the assessment cycle (Birenbaum, 2007). Steps or elements in the assessment cycle that can be classified are:

1. **Purpose or goal of the assessment**. The first step includes the determination of the educational or instructional purpose in issuing the assessment task, i.e., what is the assessment task supposed to measure, and what needs to be evidenced or shown as an outcome.

2. **Selection of assessment task**. The second step involves the composition of the assessment task itself with respect to its content or domain coverage, relative to the instruction received, and the choice of the appropriate assessment tool/instrument. That is, what content needs to be covered in the assessment task; and how will mastery or task completion be shown (choice of content and format of the assessment) (Tillema, 2010).

3. **Setting criteria for the assessment task**. Before delivery of the task, standards need to be set for qualifying task completion. This building block involves determining as well as communicating (sharing) the qualification norms. In peer assessment this element seems of special relevance since it might influence student’s trust in the appraisal (Van Gennip, Segers, & Tillema, 2009); i.e., what needs to be rated by whom on what grounds.

4. **Administering the assessment**. The actual delivery of the assessment calls for proper ways and means to offer and govern assessment task completion (time allowed, guidance offered). For example, whether students might work together on the task or prepare for different assessment moment(s); i.e., how is the assessment conducted.

5. **Scoring the assessment**. After assignment completion the actual rating and ranking of the assessment task is performed (in peer assessment by multiple raters). It involves the grading on selected criteria or standards (building block 3) done either jointly or individually, by the teacher, self and/or peers.

6. **Appraisal or “grading of the assessment”**. Next in the assessment cycle, the given scores are to be interpreted and weighted to arrive at a (overall) decision with regard to achieved levels of competence or proficiency. Evaluative arguments conducive to the arrived at decision are in order; i.e., what respect to the overall judgment to be given (McMillan, Myran, & Workman, 2002).

7. **Feedback and further promotion of learning**. Formative assessment involves feedback, as well as giving (high quality) guidance to the learner (Yorke, 2003). Therefore, in assessment for learning, the appraisal is not a matter of simply give information on passing or failing, but instead of indicating the strengths and weaknesses in achievements to provide information on how to improve further learning (James et al., 2007).

These seven building blocks cover the essential parts in assessment construction and delivery process and provide one dimension of the analytical framework; quality criteria constitute the other dimension in our review.

#### Quality criteria

In educational measurement (Brennan, 2006; Linn et al., 1991) several (construct validity) criteria have been applied to appraise assessment tasks used in the evaluation of learning and instruction. Based on previous reviews (Baartman, 2008; Birenbaum, 2007; Gulikers, Bastiaens, & Kirschner, 2004; Sluijsmans, 2002) we adopted the following set of indicators to gauge quality assurance in assessment tasks (also Van Gennip et al., 2009).

- **Authenticity**: provision of authentic assessment tasks entails measurement that is in compliance with the knowledge, skills and attitudes that will allow for (later) performance in realistic (real life) contexts (Baartman et al., 2007; Dierick & Dochy, 2001; Kerka, 2003).
1995). This criterion is often equated with active student involvement (self-directedness) to assure that assessments are perceived as relevant and engaging. We found the authenticity criterion to be linked in the reviewed literature to four, more specific criteria:

1. **Representativeness.** This criterion specifically demands the assessment task content to resemble or cover the corresponding subject matter domain (MacLelland, 2004).

2. **Meaningfulness or content fidelity.** This criterion specifically points out the importance of including worthwhile and useful task content in the appraisal of student efforts in learning (Birenbaum, 2007; Boud, 1995, 2000).

3. **Cognitive complexity.** This criterion specifically addresses the reference to higher-order skills in administering the assessment task (Natriello, 1987), in order to gauge (and stimulate) student’s cognitive abilities that scaffold their level of performance. In this sense this criterion seems especially close to the assessment for learning intent.

4. **Content coverage.** This criterion refers to breadth of coverage with respect to the instruction students received and its match with the learning process (Linn et al., 1991).

The above criteria refer mainly to the WHAT or content of the assessment task, i.e., to ensure authenticity in assignments in reference to the domain being assessed. **Transparency** as a criterion refers to assessment as a procedure (more than its content, i.e., as represented in the authenticity criterion) and points to the need of tasks to be clear, understandable, and do-able for those who have to complete them (Brown, 2004; Dochy & McDowell, 1997). A related criterion, more focused criterion) and points to the need of tasks to be clear, understandable, and do-able for those who have to complete them (Brown, 2004; Dochy & McDowell, 1997). A related criterion, more focused on and referring to the students involved in the assessment task, is **Fairness.** In generic terms, it means absence of bias to certain groups and the exclusion of irrelevant variance in the assessment task (Messick, 1980). More specifically, it says that each student should receive the equal opportunity to demonstrate the relevant knowledge, skills, abilities and competencies under appraisal (Dierick & Dochy, 2001). Transparency and fairness both indicate the need to assure clarity of the assessment for those who have to complete it. The above two criteria refer mainly to the HOW or way of administering the assessment task.

**Generalisability.** The extent to which assessment tasks and measured performances are applicable to a broader ‘universe’ of similar tasks that measure the same achievements is referred to as generalisability of measurement (Carver, 1974; Linn et al., 1991). We found this concept to be specified in the educational measurement literature under four more detailed criteria:

1. **Comparability.** This criterion is to convey that assessments must use similar conditions and criteria for everyone being evaluated to assure for consistency in appraisal (Haertel, 1991). Comparability is in a way related to fairness; therefore, we merged both criterions in our analytical framework.

2. **Reproducibility.** This criterion refers to the accurateness and consistency of standards (building block no. 2), scoring criteria (no. 3) and scoring decisions (no. 6) over time and between assessors (Brookhart, 1994).

3. **Transferability.** This criterion deals with the relevance of measurement with respect to later required performance in other situations, or in other related tasks (Linn et al., 1991).

4. **Educational consequences.** This criterion is also referred to as consequential validity of assessment (Baartman, 2008; Messick, 1994) and calls attention to (intended and unintended) impacts of the assessment on learning and instruction (Dierick & Dochy, 2001); including the so called Backwash effects (Birenbaum, 2007; Stiggins, 2002).

These last criteria of quality assurance in assessment tasks refer to the RESULT or warranty of assessment tasks (i.e., its “dependability” – Stobart, 2010).

**Linking criteria to the assessment process: an analytical framework**

In relating measurement quality criteria to different steps in the assessment cycle, we are able to establish at a more detailed level whether studies on (peer) assessment for learning adhere to or implement particular quality criteria. The purpose of this analytical framework, firstly, is to offer a tool to determine which quality criteria are taken into account regarding each step of the assessment cycle (a monitor function). And secondly, referring to its applicability, it offers opportunities for improving the quality of assessments by giving attention to each of the seven steps during assessment construction (design function). In this review, the framework will be used as a heuristic to analyze research studies on (peer) assessment to gauge plausible coverage of quality criteria in assessment for learning.

Although it might be argued that assessing the quality of (formative) assessments needs to attend to all quality criteria identified, we assume (Dierick & Dochy, 2001; Topping, 2005) that a differential attention or an uneven distribution might be present with regard to the applicability of criteria in the consecutive steps of the assessment cycle. Also, the active involvement of students in (peer) assessment might vary depending on the particular element or stage in the assessment cycle (Lizzio & Wilson, 2008). Presence of an uneven distribution of quality criteria also can be found when acknowledging the scope or range of a criterion’s applicability across the assessment cycle (its “bandwidth”). In more detail:

- **Authenticity, in essence, entails a task’s resemblance or compliance to realistic contexts to ensure meaningful learning (Gielke, Dochy, & Dierick, 2003), which in particular plays a role in determining the purpose of the assessment (step 1) and the task selection (step 2). Its subsidiary criteria: representativeness, content coverage and cognitive complexity, therefore, would then be applicable primarily to the first two steps of the assessment cycle (Sivin, 2000). Also, with regard to appraisal in assessments (step 6), meaningfulness seems an important consideration (Kerka, 1995).

- Transparency is intertwined with all the stages of the assessment cycle as is the case for Fairness. In order for tests to be fair Willingham (1999) notes that students should receive equal opportunity to demonstrate their proficiency, and scoring rules must take into account different ways of attainment.

The ‘Generalisability’ or warranty criteria: Comparability, Reproducibility, and Transferability seem especially relevant to the assessment cycle elements of scoring and appraisal (Baartman, 2008; Baartman et al., 2007) while Educational Consequence is of importance to selection of meaningful tasks and scoring criteria (Wiggins, 1993), as well as giving appropriate feedback (Freeman & Lewis, 1998).

In this sense, the ‘total’ quality management of assessment construction and administration may be secured by differently attending to relevant criteria; that is, by determining for the different steps whether the criteria relevant to each step have been taken into account (see Table 1 for an overview of expected relevancy of criteria within the assessment cycle).

In this way, the analytical framework is used to analyze our review question: to what extent are criteria as embedded in the assessment cycle, i.e., according to our proposed framework, attended to in studies on (peer) assessment to assure quality of measurement? In addition, we asked to what extent studies on (peer) assessment for learning take into account student involvement in the assessment cycle. In combining both queries we will be able to conclude, based on the studies that have applied (peer)
assessment for learning, on the degree to which assessment for learning relies on a specific set of criteria to assure its quality.

**Method**

*Selection of studies on measurement quality in peer assessment for learning*

The research databases EconLit (1990–2007), ERIC (1990–2007) and PsycINFO (1990–2007) have been systematically searched for studies implementing peer assessment. As selection criteria, the terms peer-assessment, -feedback, -evaluation, -marking and -rating were entered to gauge the abstracts. The search of research databases resulted in 1151 selected articles. Three inclusion criteria were used for selection of articles. Firstly, only empirical studies that analyzed/evaluated the implementation of peer assessment as (one of the) assessment format(s) were included. Secondly, the subjects in the studies reviewed had to be conducted in an educational context, i.e., having students as their subjects. Thirdly, since this review is intended to derive quality criteria on (peer) assessment for learning, it was decided to include only recent studies on student learning (published after 2000), to ascertain coherence and comparability in peer assessment formats. After analysis of key words, date of publishing (in abstracts), and focus on assessment for learning (in text), 132 articles which implemented (specific aspects of) peer assessment were included for further analysis.

**Method of analysis**

To review the selected studies aimed at student learning (dependent variable), the intervention variables were gauged in particular and divided studies into two categories:

(1) those who explicitly referred to quality of assessment; i.e., the quality measures used with respect to the peer assessment intervention (i.e., cycle) is clearly explicated under four queries (see below) (nine studies were found),

(2) those who were referring to (other) evaluative aspects of peer assessment, i.e., without explicit mentioning of criteria to ascertain measurement quality (for example only with indication of learning effects) (96 studies).

In an additional 27 studies no specific reference to quality of measurement was made (instead interpretive evaluations of student attitudes or perceptions on peer assessment were used). For further analysis the first type of studies are of special interest and were taken to gauge the quality criteria of the assessment. Of these, nine selected articles, one had to be discarded later on for not actually implementing a peer assessment format. Therefore, in total eight studies could be reviewed. By means of the analytical framework (abovementioned) the seven assessment cycle steps were reconstructed based on the description given in the method section of the articles. To build our analysis, four queries were performed for each step to scrutinize the given text, namely: ‘On what…’, ‘how…’, ‘when…’, and ‘by whom…’ ‘…is each step conducted’; together with an interpretive fifth query (‘why’). This last query was to gauge expressed arguments in the article for implementing peer assessment the way it was. Based on this fivefold query, the presence of a quality criterion was established. This analysis was done stepwise and replicate (along the assessment cycle dimension and along the quality criteria dimension as identified in the framework) by multiple raters.

**Selection of studies on student involvement in peer assessment for learning**

In a subsequent analysis, an analytical review was conducted to gauge the presence of student involvement for each of the respective steps in assessment cycle. A second and more generic search (no time period set) was conducted in the databases Web of Science (WSc), PsycINFO, and ERIC, with keywords ‘peer assessment’ and ‘peer evaluation’, resulting in 10,303 hits. Subsequent filters on: empirical research, education context as well as time span set to the last 20 years, which resulted in 64 articles. As a next step, articles were scanned on abstract to determine reference to quality criteria and/or student involvement. 42 articles fulfilled this sieving process and were included for more carefully reading in a qualitative review.

**Method of analysis**

The information provided in the selected studies was organized into mind maps, one for each step of the assessment construction and delivery process. Mind maps are graphic representations of relations between constructs in a two dimensional space. The software program Mindmapper 5.0 by Symtech (2008) was used to display the relatedness of criteria for our purpose. Each article’s content (statements made) was graphed according to: (a) argument by the author(s), (b) referenced conclusion based on previous research, (c) validated results from the study itself. Information in the mind maps was combined into seven narrative summaries to find the common ground on the adherence to student involvement; this was done for each step of the assessment cycle.

**Results**

**Quality criteria in assessment**

Table 2 shows an overview of the presence found for quality criteria in the selected studies. The column ‘total expected’ indicates the total amount of potentially to be found substantiations of criteria that can be attributed to assessment for learning (based on the number of studies [i.e., 8] times the number of...
<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality criteria considered in peer assessment studies.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting purpose and goals</th>
<th>Selecting tasks</th>
<th>Setting criteria</th>
<th>Administering the assessment</th>
<th>Scoring</th>
<th>Appraisal/grading</th>
<th>Guidance and feedback</th>
<th>Total found</th>
<th>Total expected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What criteria “Authenticity”</strong></td>
<td>Representativeness 3</td>
<td>Representativeness 3</td>
<td>Content coverage 3</td>
<td>Content coverage 3</td>
<td>Content coverage 4</td>
<td>Content coverage 3</td>
<td>Meaningfulness 4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>Authenticity</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Representativeness 3</td>
<td>Representativeness 3</td>
<td>Content coverage 3</td>
<td>Content coverage 3</td>
<td>Content coverage 4</td>
<td>Content coverage 3</td>
<td>Meaningfulness 4</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Meaningfulness 3</td>
<td>Meaningfulness 3</td>
<td>Meaningfulness 3</td>
<td>Meaningfulness 3</td>
<td>Meaningfulness 4</td>
<td>Meaningfulness 1</td>
<td>Meaningfulness 1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Cognitive complexity 3</td>
<td>Cognitive complexity 3</td>
<td>Cognitive complexity 4</td>
<td>Cognitive complexity 4</td>
<td>Cognitive complexity 4</td>
<td>Cognitive complexity 4</td>
<td>Cognitive complexity 4</td>
<td>32</td>
</tr>
<tr>
<td><strong>How criteria “clarity”</strong></td>
<td>Transparency 5</td>
<td>Transparency 3</td>
<td>Transparency 8</td>
<td>Transparency 6</td>
<td>Transparency 8</td>
<td>Transparency 6</td>
<td>Transparency 8</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Fairness 3</td>
<td>Fairness 7</td>
<td>Fairness 5</td>
<td>Fairness 6</td>
<td>Fairness 6</td>
<td>Fairness 6</td>
<td>Fairness 6</td>
<td>31</td>
</tr>
<tr>
<td><strong>Warranty criteria “dependability”</strong></td>
<td>Comparability of results 8</td>
<td>Reproducibility of decisions 6</td>
<td>Comparability of results 8</td>
<td>Reproducibility of decisions 6</td>
<td>Comparability of results 8</td>
<td>Reproducibility of decisions 6</td>
<td>Comparability of results 8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Transferability 1</td>
<td>Educational consequences 1</td>
<td>Educational consequences 1</td>
<td>Educational consequences 1</td>
<td>Educational consequences 1</td>
<td>Educational consequences 1</td>
<td>Educational consequences 1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Educational consequences 4</td>
<td>Educational consequences 3</td>
<td>Educational consequences 2</td>
<td>Educational consequences 3</td>
<td>Educational consequences 3</td>
<td>Educational consequences 3</td>
<td>Educational consequences 3</td>
<td>129</td>
</tr>
</tbody>
</table>

Total found 22 15 23 11 31 19 7 129
Total expected 48 48 32 16 32 32 32 272
relevant or applicable steps, i.e., based on Table 1. This total is contrasted with the actual reference to a specific quality criterion in the reviewed studies (leading to a total coverage ratio of .48) (Table 2).

With regard to the first step in the assessment cycle (purposes and goals) authenticity (WHAT) criteria as well as clarity (HOW) criteria (fairness and transparency of the assessment) are being considered in the studies, but to a different degree. Only in one study (Dannefer et al., 2005) all criteria were noted. Educational Consequences with regard to the purpose of assessment was taken into consideration by four studies. For instance, the study by Langan et al. (2005) stated that “peer assessment provides a useful mechanism to develop many positive qualities in students studying in higher education”, thereby indicating the purpose of peer assessment to enhance learning.

With respect to the second step, constructing assessment tasks, the selected studies seem to pay attention to authenticity criteria predominantly. For example, cognitive complexity is considered when Davies (2004) mentions that awarding a ‘mark for marking’ is to be supportive to demonstration of higher-order skills in assessment. Surprisingly, fairness was not considered in the studies as related to assessment task construction.

In analyzing the third step choosing scoring criteria, it turned out that fairness and transparency are explicitly addressed. Also half of the studies paid attention to content coverage in scoring criteria. For instance, the study of Cho, Schunn, and Wilson (2006) explicitly addresses the specific dimensions of the task to be evaluated/scored, by describing anchor points that cover the relevant aspects for peers in reviewing a task.

In the step: administering the assessment almost no explicit quality information was found in the studies on its execution. Fairness is considered to some extent and is usually operationalised as the proper carrying out of assessment process. Only some studies mention that scoring rules should be equal for all students. Orsmond, Merry, and Reiling (1996) for example, describes assessment implementation in which the scoring instructions need to be the same for all students.

In comparison, with regard to scoring studies take into account many of the identified quality criteria outlined in the framework (31 out of the total 129). Especially, the reviewed studies consider the criteria of transparency, fairness; and, for the first time we noted as well, the criterion of generalisability, in the sense of comparability of results. The study of Ryan, Marshall, Porter, and Jia (2007) is a clear example on how researchers operationalise this criterion. In their study, students were informed of the scoring policies (transparency); the assessment was used also to rank classroom participation (comparability of results). An operationalisation of fairness was to ascertain that markings are private and confidential (Willingham, 1999).

With regard to appraisal, we found that authenticity criteria (meaningfulness and content coverage of the assessment) as well as transferability (breadth of coverage) were not considered in the studies. Transparency, fairness and reproducibility of decisions were considered in six of the eight studies. In the Bastick (1999) study for example, these criteria were considered in a transparent scoring procedure for deriving individual marks. The assessment was a weighted appraisal by students who combined their confidential peer assessments to arrive at individual marks. This procedure made the interpretation reproducible as well as fair.

Notably, studies gave least consideration to quality criteria with regard to providing guidance and feedback (7 out of the potential 129). Only two studies pay attention to this step. Cho et al. (2006) and Dannefer et al. (2005) consider this step by giving students meaningful feedback or opportunity to improve. In the Dannefer et al. (2005) study for instance students are informed to bring their report to a meeting and discuss how the feedback might contribute to constructing an individualized learning plan (Educational Consequences).

Student involvement in assessment

Seven narrative summaries were constructed based on the mind maps to represent the coverage and relevancy found for student involvement as a quality criterion with regard to each step of the assessment cycle. An example of a mind map is added in Appendix A.

Step 1: Assessment purpose

Several studies indicate (Rust, Price, & O’Donovan, 2003; Sluijsmans, Dochy, & Moerkerke, 1999) that setting purpose and goals of assessments needs to involve students, since without explaining the rationale of an assessment it is hard to encompass its relevance (Bloxham & West, 2007). Sadler (1989) points out the importance of ownership to make goals more meaningful and transparent. Transparent goals mean that the assessment language needs to be clear to students (Liu & Carless, 2006). Also fairness is mentioned as an important criterion for students to be involved in (Lizzio & Wilson, 2008; Poulos & Mahony, 2008).

Step 2: Selecting tasks

It is mentioned frequently that students have sceptical attitudes toward (peer) assessment (Papinczak, Young, & Groves, 2007). To make tasks meaningful McDowell (1995) suggests to establish a linkage between the task and personal interests of students. The assessment tasks have to be embedded in and be part of the regular courses to optimize peer involvement (Liu & Carless, 2006). Sivan (2000) reported that students also do positively accept the assessment method, when they think it contributes to their learning. Also Falchikov (2005) reported positive student attitudes. Fairness in the choice of tasks seems to be a most significant criterion for students. Cheng and Warren (1997) conclude that peer assessment was considered more fair than assessments in which students were not allowed to participate in its construction.

Step 3: Setting criteria (see Appendix A for a mind map)

Students find it difficult to use scoring systems and understand scoring criteria (Bloxham & West, 2007; Carless, 2006; Dochy et al., 1999; Papinczak et al., 2007). This seems due to a lack of transparency of criteria (Carless, 2006; Papinczak et al., 2007), causing students not to recognise the criteria and therefore probably avoid using them (Orsmond et al., 1996). Students are concerned with the possibility of unfair and inaccurate marking (Vu & Dall’Alba, 2007). Liu and Carless (2006) bring forward that understanding the criteria is fundamental to collaborative forms of assessment, like peer assessment. Training on how to use scoring criteria seems a solution. It results in more effective peer assessment (Xiao & Lucking, 2008), more reliable scoring (Falchikov & Goldfinch, 2000; Topping, 2009; Zhang, Johnston, & Kilic, 2008); and deep understanding, and sense of ownership (Falchikov & Goldfinch, 2000). According to O’Donovan, Price, and Rust (2004), training with a structured approach yielded greater understanding and explicit articulation of assessment criteria (McMillan, 2007; Sivan, 2000). The use of exemplars can help students to understand how to use criteria in assigning marks against standards (Vu & Dall’Alba, 2007).

Step 4: Administering the assessment

Xiao and Lucking’s study (2008) found that providing as well as receiving peer assessment was worthwhile and beneficial to students. The quality of peer assessment is influenced by student conceptions and their previous experiences (Vu & Dall’Alba, 2007). Sivan (2000) concluded that discussing with students issues of bias and fairness as well as misusing the assessment is beneficial. Discussion is also important for transparency as a way for students to understand the administering of assessments (Ryan et al., 2007), and their role in it (Sluijsmans et al., 1999). But, as Liu and Carless (2006) bring forward, probably all actions to increase transparency of the peer assessment process will be beneficial.
Step 5: Scoring. Students lack confidence in their own assessment abilities, and express concerns about their capability to mark fairly (Falchikov, 1995; McDowell, 1995). Less than half of the students thought they were able to mark fair and responsibly (Cheng & Warren, 1997). Also Orsmond et al. (1996) report that the students feel 'unqualified' to assess peers and were even reluctant to do it. Struyven, Dochy, and Janssens (2005, p. 334) talked about "mixed feelings about being capable of assessing each other fairly". This indicates a lack of confidence in ability. Biggest threat to fairness in peer assessment is bias due to friendship marking and lack of honesty (Papinczak et al., 2007); especially when there is no training in peer rating (Dancer & Dancer, 1992), or when they perceive the environment as not safe for interpersonal risk-taking (Van Gennip et al., 2009). Students are reserved to be critical, in order not to cause friction (Vu & Dall’Alba, 2007). Students do not seem to have many concerns about the accuracy of the assessment (Falchikov, 2005). 80% of the students indicated that the received assessment was accurate (Xiao & Lucking, 2008). The assessment and the comparability of the results can be improved by practicing the particular assessment task (Cheng & Warren, 1997).

Step 6: Appraisal. Peer appraisals can be given in the form of a grade, or directions for further learning. In this step fairness is important to students (Lizzio & Wilson, 2008; Poulos & Mahoney, 2008; Struyven et al., 2005). Students are concerned with their assessment outcomes and grades (Lizzio & Wilson, 2008). Topping (2009) notices that the accuracy of peer assessment can be questioned.

Step 7: Guidance and feedback. Students find it important to receive meaningful feedback, and demonstrate higher levels of satisfaction with qualitative feedback (Topping, 2009; Xiao & Lucking, 2008); they prefer specific feedback above general remarks (Poulos & Mahoney, 2008). The use of exemplars can contribute to more meaningful feedback (Orsmond et al., 1996). Sadler (1989) notes that students should be trained to interpret feedback in order to make connections between the received feedback and their own work. The lack of useful feedback is seen as a problem by students (Carless, 2006; Topping, Smith, Swanson, & Elliot, 2000); implicating that Educational Consequences of the feedback are often limited. Lizzio and Wilson (2008) bring forward that the level of respect afforded in the feedback may be an important factor (i.e. the tone of the comments) and should be provided in a language that the students already know and understand. That is, the students have to understand the feedback, before they can make use of it.

Discussion

This study was carried out to detect quality criteria that are attended to in constructing and using peer assessment for learning. The analysis was intended to substantiate a framework which relates quality criteria of assessment for learning to seven steps in the assessment (construction) cycle. Such a framework was considered to scaffold the development and design of (peer) assessment tasks used to enhance learning (Ploegh et al., 2009).

Our findings show that specific quality criteria are taken into account in relation to precise steps in the assessment cycle, while others have a more generic deployment. Firstly, quality criteria in first steps of the assessment cycle (i.e., setting assessment goals and selecting assessment tasks) were specifically related to representativeness, content coverage and content fidelity while these criteria were less prominent in later steps. As a generic criterion meaningfulness seems to apply to almost each individual step in the assessment cycle. Secondly, the criteria fairness and transparancy were considered applicable to all steps of the assessment cycle. Thirdly, with regard to the Generalisability criteria, we see a divide between Educational Consequences and the other forms, in that the former is regarded in almost all steps while the latter are being reserved for the final stages in the assessment cycle. This might indicate a generic relation between Educational Consequences and the assessment’s warranty as a whole.

In conclusion, it seems that criteria use in quality assurance is mainly arranged to produce clear and meaningful assessment tasks which can warrant instructional implications for further learning. Based on our second review specifically which addressed student engagement in the assessment it can be concluded that each step in the assessment cycle presents itself as an opportunity to involve students in their own evaluation/grading process (Falchikov, 2004). This interpretation is supported by the finding that most quality criteria were taken into consideration in the step of scoring.

Furthermore, with regard to the evaluation of the framework, it was noted that use of quality criteria in the reviewed studies was considered predominantly in case of goal selection, selecting tasks, setting criteria, and scoring (steps 1–3 and 5), and not typically for appraisal and feedback (i.e., the later steps in the assessment cycle – steps 4–7). Considering the aim of assessment for learning, it is surprising that criteria in the step guidance and feedback were measured the least (Topping, 1998). It is also noteworthy that the criterion transferability is hardly taken into account in the studies reviewed. It is possible that transferability is embedded in other quality criteria (notably representativeness or content coverage) and, therefore, maybe (or should be) considered important, as well when constructing the assessment task. In this respect, we may conclude that some criteria in the framework seem to overlap and tend to be taken into consideration jointly (e.g., transparency, fairness and comparability of results/reproducibility of decision in case of the last two steps). Other quality criteria, such as meaningfulness, may include different notions of authenticity and showed a broader ‘bandwidth’ of applicability than initially posited (Table 1), while others probably still need more operational refinement to become specific an detailed enough to be handled for a specific step in the assessment cycle. The generic applicability of fairness and transparency might imply a certain indifference to any of the steps in the assessment cycle. We also found, based on a comparison of studies, that various operationalizations of the same criterion were given in the studies. For example, fairness is operationalised as having equal procedures for all students, or as averaging peer and tutor marks, or as using anonymous and confidential markings. We also found considerable variation in detail with which quality criteria were specified; of the 134 studies only eight were detailed and specific on their measurement of peer assessment.

Our review of student involvement in assessment adds to our observation that clarity (i.e., fairness) and meaningfulness are considered main criteria in the construction and administration of assessments for learning. We found that peers as assessors feel they need clear criteria to meaningfully appraise each others work. Students often are involved only in the steps of scoring and appraisal (i.e., not in the preceding step of goal selection or subsequent steps of feedback giving). The reviewed studies point out that students find it difficult to use scoring rules per se (O’Donovan, Price, and Rust, 2008). Lack of understanding (of goals and purpose) increases the possibility of unfair and non-transparent marking, however (Vu & Dall’Alba, 2007). Besides, students express fear at exposing their marks to peers (McInerney, Brown, & Liem, 2009). Therefore, understanding of the criteria by students is fundamental for formative assessment to be fruitful. Student training in criteria may contribute to the effect of more successful peer assessment as well as improved learning (Magzoub, Schmidt, Dolmans, & Abdelhameed, 1998). From an assessment for learning perspective, it is important as well that student can be involved in providing explicit feedback to each other. Involvement in criteria selection and goals selection is an effective way to develop understanding and responsibility for their learning (Sivan, 2000).
In conclusion, our review of measurement criteria and student involvement in (peer) assessment demonstrates that the quality criteria for assessment are unevenly distributed and differently paid attention to in the different steps or building blocks of the assessment cycle. Content fidelity and clarity seem to be taken into account most widely in current studies implementing peer assessment. Quality criteria also are applied mostly to scoring issues of the assessment (and less with regard to feedback provision and guidance for further learning). Even though we found that most studies in our review used quality criteria implicitly, it still implies they are considered as important to assessment for learning. However, it would scaffold both transparency and meaningfulness of the assessments if, in the future, these quality criteria will be taken into account more explicitly (not only in research; McNerney et al., 2009; Sluijsmans, Moerkerke, van Merriënboer, & Dochy, 2001), so as to manage the construction and administration of assessments by means of a clear framework on quality criteria. Such a more detailed implementation of criteria could then optimize the measurement quality of assessments in utilizing the different steps in the assessment cycle. With such a framework, it will be possible to determine to what degree certain criteria need to be taken into account, to give an overall ‘quality judgment’ of the assessment.

Use of the framework in linking assessment construction steps to quality criteria might also offer opportunities to detect and develop more specific guidelines to refine and successfully implement assessment tasks and lead to clear and specific instructions on quality criteria. Our findings support the idea that in order to establish quality of assessment for learning, it is important to explicitly regard quality criteria relative to the designated steps in the construction of an assessment (Torrance & Pryor, 2001). In this respect, further development and specification of a robust framework on quality criteria with a corresponding conceptual refinement of its operationalizations is needed in order to increase its relevance and warranty in assessment for learning.

Acknowledgements

We would like to thank Karin Ploegh for her contribution in thinking and writing about the conceptual underpinning of this paper and her participation in important parts of the data collection process.

Appendix A. Mind map of scoring criteria of the assessment task
References


