Competence Assessment as Learner Support in Education

Chapter 28
Competence Assessment as Learner Support in Education

Cees van der Vleuten, Dominique Sluijsmans, and Desiree Joosten-ten Brinke

28.1 Introduction

The assessment of professional competence has developed progressively in the last decades following the changes occurring in education. Education has shifted from an input model of education to an outcome-based model of education (Chappell et al. 2000). Instead of requiring certain hours in a curriculum on certain disciplines (the input model), modern education programmes are based on a defined set of outcomes or competencies (the output model). All courses and the assessment are then aligned to these outcomes. A second major shift is that many of these outcomes or competencies move beyond the knowledge domain, into more authentic professional skills or general competencies relevant for success in the labour market (Semeijn et al. 2006). Being able to work in a team, being able to communicate, being able to write academically and being able to behave professionally are examples of these general competencies. They are less domain specific, hence their general or generic nature. Both success and failure in the labour market are associated with these kinds of skills (Heijke and Meng 2006). As a result modern curricula pay more attention to the development of these skills. Finally, a third major change is a didactical one, where education is moving from atomistic to holistic learning, from teacher-centred learning to student-centred learning, from an exclusive focus on

C. van der Vleuten (✉)
Maastricht University, Maastricht, Netherlands
e-mail: c.vandervleuten@maastrichtuniversity.com

D. Sluijsmans
Zuyd University of Applied Sciences, Heerlen, Netherlands

D. Joosten-ten Brinke
Open University in the Netherlands, Heerlen, The Netherlands

Fontys University of Applied Sciences, Eindhoven, The Netherlands

M. Mulder (ed.), *Competence-based Vocational and Professional Education*, Technical and Vocational Education and Training: Issues, Concerns and Prospects 23, DOI 10.1007/978-3-319-41713-4_28

lecturing to more active learning methods and from highly teacher-led structured learning to self-directed learning (Merrienboer and Kirschner 2007). The assessment of professional competence has followed these educational developments. Therefore, assessment has moved beyond the knowledge domain, towards more complex assessment of skills in authentic contexts. In the literature, a difference is made to testing and assessment (Segers et al. 2003). Testing would refer to more classic approach of standardized testing of mainly cognitive functions, whereas assessment includes more modern forms and more authentic forms of assessment. We prefer the term assessment as an overarching term for all forms of assessment in which abilities of learners are assessed, whether conventional or modern and innovative and whether yielding quantitative or qualitative information. In this chapter we will use three different perspectives on assessment to illustrate the implications of educational developments for assessment practice: assessment *of* learning, assessment *for* learning and assessment *as* learning.

In an assessment *of* learning perspective, the focus is on optimizing appropriate decisions on our learners (Segers et al. 2003). This is associated in the literature with the term summative assessment. Have our learners achieved certain educational standards? Can we account for taking the right decisions over them? The emphasis is on the credibility of the decision-making function of assessment. In the assessment *of* learning perspective, an overview will be given on various classes of assessment methods, each assessing competence with different levels of authenticity. Subsequently, lessons learned are summarized from the research that has emerged from these classes of assessment methods. This will cover issues of reliability, validity, objectivity and impact on learning. The overview will lead to a set of general recommendations for assessment. One very central recommendation will be that any individual assessment method will always be a compromise on its qualities and will have serious limitations.

In an assessment *for* learning perspective, the focus of assessment is on its effect on learning (Black and Wiliam 2009). Does the assessment provide meaningful feedback to learning? How does the assessment support the on-going learning process? How may assessment promote deeper learning strategies or certain developmental outcomes? We will provide an overview of the literature on assessment *for* learning, also named formative assessment. More particularly, an overview will be given of the most current methods used for formative assessment in the classroom and the requirements for being effective for learning. The overview is informative for teachers wishing to use assessment to guide learning.

Finally in an assessment *as* learning perspective, both the decision function and the learning function are united in one single synthetic approach to assessment (Clark 2010). Assessments are seen as an integral approach by looking at the design of full assessment programmes. Limitations of individual assessment methods can be overcome by looking at assessment from a programmatic approach. Comparable to a curriculum, assessment programmes are planned, implemented, governed, evaluated and adapted. We will discuss an approach to programmatic assessment in which the perspectives of assessment of learning and assessment for learning will

be intertwined. Finally, an illustration is given of an existing assessment practice that was designed programmatically.

In the following, each assessment perspective is outlined. We will refrain from providing methodological, theoretical or psychometric overviews in assessment. Methodologically oriented theories of assessment are well developed and abundantly available (Zwick 2006; Kane 2001). Our intent is to provide assessment insights on each perspective that may have direct relevance to educational practice.

28.2 Perspective 1: Assessment of Learning

28.2.1 Classifying Methods of Assessment

In any method of assessment, we can make a distinction in the stimulus format and the response format. The stimulus format is the kind of task one gives to the person being assessed. A stimulus might be asking for a fact or it might be a rich scenario. The response format is the way one captures the response. This could be a small menu of options, a write-in response, a long essay, an oral form, etc. The number of assessment methods resulting from combinations of stimulus and response formats is really infinite and impossible to reproduce and explain. Instead we will use a very simple model of competence to classify methods of assessment. Miller proposed a pyramid of competence with four different layers (Miller 1990). Figure 28.1 gives an overview of the layers of the pyramid and the way these layers can be assessed. At the bottom of the pyramid is "knows" representing the assessment of factual knowledge. This is mostly done through written or computer-based tests testing for facts. Methods may range from multiple-choice tests, short answer tests or oral

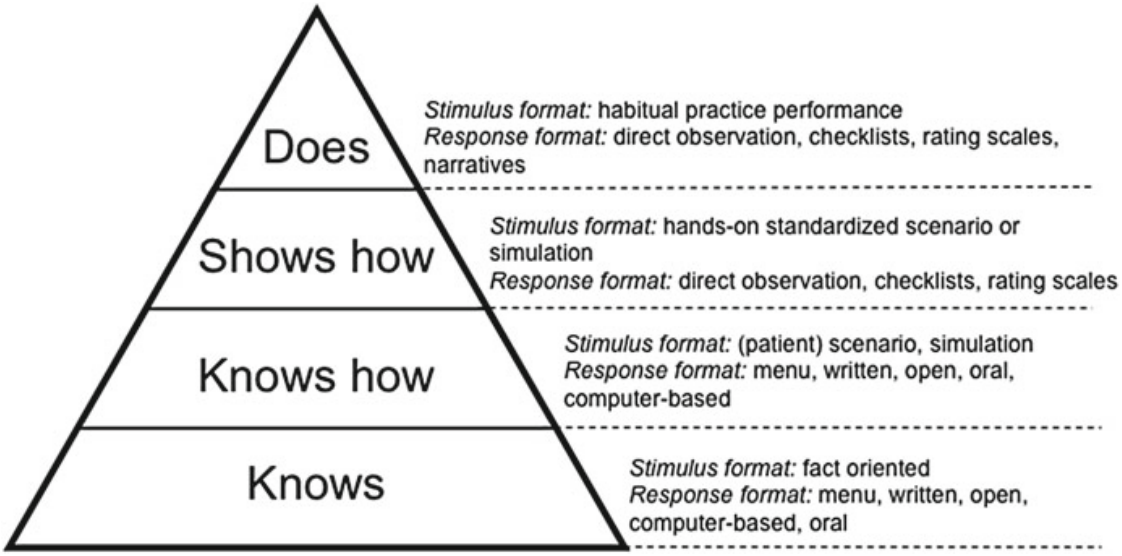

Fig. 28.1 Miller's pyramid of competence with each layer classifying methods of assessment

examination testing for facts. Once a learner is able to use the knowledge and apply it to cases or is able to solve or reason through problems, this is then called 'knowing how'. Methods of assessment may be very similar to the first level in terms of the response format, but now they typically rely on richer stimulus formats such as cases, quizzes, problems or scenarios. Higher-order cognitive skills are assessed, in which reasoning and application of knowledge is required. Open-book exams fall into this category too (Koutselini Ioannidou 1997). Other examples are essays, oral assessments where learners present cases or longer reports on project work or research projects (Segers and Dochy 2001).

Where the first two levels assess cognitive skills, the top two layers assess behavioural performance. At the 'shows how' level, behaviour is assessed through a simulation of professional tasks that is being judged on checklists, rating scales or in narrative forms. A widely known example is the assessment centre. Candidates are presented with tasks in a simulated work setting and have to deal with these assignments professionally. Assessment centres are used in many different contexts such as in career decisions in the police force (Feltham 1988), selection of naval officers (Jones et al. 1991) and in business (Sagie and Magnezy 1997). In health sciences such as medicine, dentistry and nursing, they use what is called Objective Structured Clinical Examinations (OSCEs) (Petrusa 2002). An OSCE consists of several stations with each station presenting a different simulation. Candidates rotate through these stations and a different examiner in each station scores the performance using a checklist or a rating scale. In a number of countries in the world, these performance tests are part of national licencing examinations for medical students and are used on a very wide scale in virtually all schools training health professionals. The next level of competence, the 'does' level is when actual behavioural performance is being assessed in real-life work settings. Work-based assessment can be assessed indirectly, for example, by judging artefacts or products of the work, or directly by judging observed performance. For example, an experienced teacher may observe a student teacher while teaching in a classroom. Other typical instruments are ratings by supervisors or ratings by multiple persons such as in the multisource feedback (MSF) or 360° feedback. In the latter, a set of ratings is used from a range of relevant co-workers or clients or other relevant persons, providing quantitative (rating scales) and qualitative information (narrative information) on the person being assessed (Brett and Atwater 2001; Wood et al. 2006). This information is reported back to the learner in an aggregated form and often complemented with a self-assessment. Another instrument that is rapidly gaining interest is the portfolio. In a portfolio the burden of evidence is reversed from teacher to learner. Based on preset criteria, usually derived from output definitions or competency definitions, the learner has to demonstrate fulfilling these criteria through presenting evidence (artefacts, other assessment information, recorded activities) and by reflecting on the evidence. Portfolios have gained tremendous popularity in virtually all fields of education and several reviews on their value have been written (Butler 2007; Van Tartwijk and Driessen 2009). What is important to note is that the first three layers of Miller's pyramid are about standardized testing technology, whereas the fourth layer is on non-standardized assessment technology. In standardized assessment, all

assessment conditions are as much standardized as possible for all test takers, even the more authentic assessments in the third layer. When assessing in real practice, standardization is impossible and assessment is based on more subjective, often more holistic judgements in non-standardized conditions.

28.2.2 Principles of Assessment

Each of the layers of Fig. 28.1, in fact each of the instruments in each layer, could be a full chapter, so it is impossible to be comprehensive here. Instead we are taking a different meta-analytical approach and sketch the consistencies found in the research that lead to principles of assessment of a more generic nature applying to more than one or all assessment methods. We partly rely on the work of Van der Vleuten et al. who sketched a number of these consistencies (Van der Vleuten et al. 2010). Table 28.1 summarizes a set of slightly modified principles for standardized and non-standardized assessment, respectively, and we added practical consequences for its implications to educational practice. We will discuss Table 28.1 systematically.

One very consistent finding is that competence is *specific and not generic*. Whatever is being measured and whatever method it is assessed with, competence and resulting performance are contextually bound. If we sample one element of performance (in an item, or question, situation, station, scenario), that performance is not very predictive for performance on another element. Therefore, many elements need to be sampled in order to be able to make an inference on someone's ability independent of the sample used in the assessment. This is often referred to as the reliability of an assessment or the reproducibility of the findings. The phenomenon has been termed differently in different domains as content specificity or task variability (Swanson et al. 1995; Shavelson et al. 1993). Two elements in any assessment 'disagree' considerably about the performance of a learner, and therefore we need many elements in any assessment. How many elements are needed varies from method to method and from situation to situation. By looking at overall testing time from various methods across Miller's pyramid at least 3 to 4 h of testing time is required to produce stable findings (Van der Vleuten and Schuwirth 2005). Therefore, shorter tests will deliver quite a few false positive and negative decisions. The practical implication is that we should sample contexts or test elements as much as possible within one particular assessment. For example, in an assessment centre, one cannot rely on sampling a few situations. In order to make a confident decision on the competence of an individual, we need many situations, often requiring more than 3 or 4 h of testing time in total. We should therefore be careful with one moment of assessment. We should preferably combine information from different assessments or from assessments over time. One measure is really no measure, at least when pass/fail decisions need to be taken that have serious consequences for the learner.

At the same time, it has been found that content specificity equally affects all method of assessment, so one instrument is not really inherently better than another

Table 28.1 Consistencies in assessment research and their practical implications

Standardized assessment Assessing 'knows', 'knows how', 'shows how'	Practical implications
Competence is context specific, not generic	Broadly sample performance across content within each assessment
	Combine information across assessment or across time
	Avoid high-stake decision on a single assessment
Objectivity is not the same as reliability	Use holistic professional judgement when it is needed
	Use many subjective judgements in combination
What is being measured is more determined by the stimulus format than by the response format	Any method may assess higher-order skills
	Produce stimulus formats that are as authentic as possible (scenarios, cases, etc.)
Validity can be 'built-in'	Organize quality assurance cycles in item and test development
	Use peer review
	Use psychometric information
	Use student input
Unstandardized assessment Assessing 'does'	Practical implications
Bias is an inherent characteristic of professional judgement	Use sampling to reduce systematic errors
	Use procedural measures to reduce unsystematic errors that build to the credibility of the judgement
Validity lies in the users of the instruments, more than in the instruments	Prepare and train assessors and learners for their role in the assessment
	Create working conditions that embed assessment possibilities
Qualitative, narrative information carries a lot of weight	Use words for assessing complex skills
	Be aware of unwanted side effects of quantified information
Feedback use needs educational scaffolding	Create feedback dialogues
	Create feedback follow-up
	Create meaningful relations between teacher and learners
Overall	Practical implications
No single method is perfect	Vary in use of assessment methods
	Combine information from multiple assessment sources

instrument in terms of reliability (Van der Vleuten et al. 1991). Methods of assessment that are traditionally considered to be more objective do just as well (or just as poor) as more subjective methods, all depending on the sampling. Objective assessments (e.g. an MCQ exam) can be unreliable when it samples only a few items, while more subjective formats (e.g. an oral examination) can be reliable when it uses sufficient samples of content and assessors. Usually in standardized assessment, two assessment contexts 'disagree' more on the ability of a learner than two assessors. Therefore, when multiple contexts are being used for wider sampling and different assessors are used for the different contexts, reliable scores and decisions can be achieved. Yet in other words, *reliability is not the same as objectivity* and should not be confused. It implies that whenever professional judgement is inherently required, we should use it and not avoid it. Professional judgement is needed to assess complex skills (e.g. poetry writing, writing a scientific text) or complex performances (e.g. dealing with a client, communication in a team, a musical performance, a surgical intervention, etc.). Professional judgement may come from an expert but also from peers or co-workers. Sometimes the professional judgement comes from the self.

Sometimes teachers make claims on the virtues of certain methods as if methods of assessment have a fixed inherent validity. However, *what is being measured is more determined by the stimulus format than by the response format* (Schuwirth and van der Vleuten 2004). The task in the assessment that is given to the learner defines much more what is being measured than the way we capture the response. So, for example, open-ended question may assess factual recall (Miller level 1) and a multiple-choice item may assess problem solving (Miller level 2), all depending on what is being asked in the question itself. Whether the response is a short menu (like in the MCQ) or an open answer has relatively little impact on what is being assessed. The practical implication is that we should be more mindful about the stimulus format rather than the response format. In educational practice, often the reverse is the case and choices are often based on naïve assumptions about inherent qualities of an assessment method. Authentic tasks are typically used for assessing higher-order skills. The way these authentic tasks are written and presented to the learner is really important for being able to assess these higher-order skills. They can be operationalized in virtually every response format.

Many in-house assessments in educational practice suffer from quality (Jozefowicz et al. 2002). Quality assurance around item and test development has a dramatic effect on the quality of the assessment (Verhoeven et al. 1999). Therefore, much of the *validity can be built into the assessment*. Quality assurance can be done before and after a test administration. Measures before test administration may include peer reviewing of test material; piloting test material; developing scoring standards, checklists or scoring rubrics; training of assessors; and choosing an appropriate standard setting method (Cizek 2001). After test administration, item and test statistics may be used for deciding to drop poor performing items. Student comments may be quite useful for that purpose as well. When psychometric expertise is available, tests may be corrected for difficulty variations and scores of items and students expressed on a similar standardized scale. All these measures have a

tremendous effect on the quality of the assessment. Naturally, they are also resource intensive. However, assessments without any quality control, as it is often the case in educational practice, risk poor quality leading to invalid assessment. High-quality test material is expensive. One resource saving strategy is to share test material across institutions within domains. Within the medical domain, for example, assessment alliances are created that share written and performance-based assessments across member institutions, both nationally (www.medschools.ac.uk) and internationally (www.idealmed.org). Unfortunately, cross-institutional collaborations occur rarely in educational practice.

That *assessment drives learning* has been documented since very long time (Frederiksen 1984). A distinction is made in pre-, post- and pure learning effects (Dochy et al. 2007). A learner will prepare for an assessment (pre-learning effect) or may learn from the feedback of the assessment (post-learning effect). Experimental research has shown a consistent effect that when instruction is paired with assessment, learner performance is improved (Karpicke and Roediger 2008). This is the pure learning effect also called the testing effect. It has proven to be a robust effect across a range of different educational settings (McDaniel et al. 2007) and different methods of assessment (Agarwal et al. 2008), including performance assessment (Kromann et al. 2009). Pre-learning effects can be very diverse and often promoting poor learning strategies (Cilliers et al. 2012). Providing feedback to promote a post-learning effect may not always be effective. This will be further discussed in the section on the assessment for learning perspective.

Unstandardized assessment directly assessing performance in work settings (the stimulus format) completely relies on professional judgement of assessors having observed the learner. This is usually captured by holistic judgements using generic rating scales, questionnaires or narrative comments (the response format). As soon as holistic judgement is used, bias is introduced. For example, the 'failure to fail' is a well-known problem in performance assessment (Dudek et al. 2005). Assessors tend not to use the lower end of a rating scale, probably due to adverse consequences of a negative judgement. This is the 'leniency effect' in performance assessment. There are many other rater effects and those are probably all in operation when a holistic performance judgement is made (Murphy and Cleveland 1995). Assessors are not passive instruments, who can be easily calibrated to represent a 'true score' as assumed in psychometric theory (Govaerts et al. 2007). Their judgement will be influenced by their personal experience and expertise, by the work context and by the relationship with the person being assessed (Berendonk et al. 2013). *Therefore*, *bias is a natural given in any form of judgement*. The classic response to the reduction of bias is to 'harness' the judgement by making it more analytical, for example, by standardizing the performance task, by formulating strict performance criteria operationalized in checklists, followed by training programmes trying to calibrate the assessors. Performance observation in real practice is by definition not standardized. In the same vein, complex skills (e.g. communication, collaboration, professionalism, leadership) are extremely difficult to define in atomistic elements. Attempts to such definitions lead to reductionist representations of the construct being assessed and when implemented in assessment practices then

lead to trivial performance by learners complying with the strict criteria. In habitual performance assessment, we need other strategies to deal with the natural inherent bias. We suggest two approaches.

The first applies the first principle discussed and that is sampling. By sampling a number of observations from different assessors, variation between assessors will average out. This is confirmed by considerable research. Several studies based on large data sets indicate that very reasonable performance samples lead to sufficient reliability (Wilkinson et al. 2008; Moonen-van Loon et al. 2013). For example, if we assess the performance of a learner in a practice setting, we need some eight observations (Van der Vleuten and Schuwirth 2005). Similarly, if we assess the performance of a learner through a multisource feedback, somewhere between six and ten raters may suffice (Wood et al. 2006). Sampling may therefore be an important strategy for reducing assessor bias.

A second strategy is of a completely different nature and applies particularly when information is aggregated across different (assessment) sources into a high-stake pass/fail or promotion decision. Driessen et al. (2005) suggest using procedural strategies that are derived from qualitative research to build rigour to the assessment. An example may clarify this. When a portfolio is used for aggregating relevant information on the performance of a learner, that portfolio needs to be judged. It is possible to think of many procedural measures that bring credibility or trustworthiness to a decision made over the quality of the portfolio (qualitative research strategy in brackets):

- Having a committee of (independent) experts (stepwise replication)
- Increasing size of the committee (stepwise replication)
- Tailored increased volume of expert judgement and deliberation proportional to uncertainty of information at hand (triangulation)
- Justification of decisions (thick description)
- Appeal procedures (audit)
- Previous feedback cycles (making decisions unsurprising) (prolonged engagement)
- Incorporation of learner view (member checking)
- Training of examiners (prolonged engagement)
- Scrutiny of committee inconsistencies (structural coherence)

In this way, an essentially subjective professional judgement is fortified alternatively, not by standardization neither by objectification but by due process measures. By doing so the holistic nature of what is being assessed is maintained, and trivialization of the learning processes to achieve these complex skills is avoided. In essence, what is important is *not* to ban subjective judgements from the assessment process as is done traditionally but to use professional judgement whenever they are appropriate. Professional judgement is the core of expertise in many professional domains, so should it be in assessment as well whenever it is appropriate.

In unstandardized assessment, formative and summative functions are typically combined, and feedback is given to the learner (Norcini and Burch 2007). Therefore, the role of the interaction between assessor and learner is crucial for the success of

unstandardized assessment. The way feedback is given and received really determines the value and the quality of the assessment. Unlike standardized assessment where we enhance quality through quality assurance (we can 'sharpen' the instrument), in unstandardized assessment *the users of the instrument are eminently important* (so we have to 'sharpen' the people involved). Both feedback giving and feedback reception are skills that need to be developed. In a study on the success of implementing work-based assessment on a large scale in postgraduate medical training, buy-in from the supervisors was key to the success of it (Fokkema et al. 2013). When the people are important, then we need to invest in them. Capacity building becomes important. Teachers, supervisors, co-workers and learners themselves need to be prepared and trained for their role in the assessment. The purpose is less to standardize or to calibrate assessors but to make them effective feedback givers. Teaching and assessment roles become very intertwined.

A famous quote by Einstein states: *Not everything that counts can be counted, and not everything that can be counted counts*. Assessment is traditionally associated with quantification, with numbers, scores and grades. Complex skills and behaviours, however, are very difficult to quantify, and when so they provide little information for feedback and remediation (Shute 2008). Research shows that learners appreciate narrative feedback because it is more useful to them (Govaerts et al. 2005; Ellis et al. 2008; Pelgrim et al. 2011). In performance assessment therefore *narrative information carries a lot weight*. Narrative information provides the nuance that numbers are not able to provide. Putting a metric on something that is hard to capture in a metric may again lead to reductionist practices, for example, students hunting for grades, not accepting lower grades or adherence to the bare minimum only. The implication is that we should invite assessors to use narratives, to describe behaviours in their feedback whenever that is appropriate. With performance assessment, we slowly move from numbers to words, a movement that may have quite some implications (Govaerts and Van der Vleuten 2013).

Research indicates that the provision of feedback does not guarantee its use (Hattie and Timperley 2007; Harrison et al. 2013). A lot of feedback is simply ignored (Hattie and Timperley 2007), all depending on the kind of feedback (Shute 2008), its credibility and the culture in which it is given (Watling et al. 2013; Harrison et al. 2014). Emotions also play an important role (Sargeant et al. 2008) as well as reflection on the feedback (Sargeant et al. 2009). We may conclude that *effective feedback needs to be scaffolded* by educational arrangements in which reflection and follow-up occur. One way of scaffolding is to create dialogues around feedback in social interactions. This can be done in peer groups or by mentoring. Building entrusted relationships, such as in mentoring, in which feedback and follow-up are discussed with learners is a very powerful approach to feedback use (Driessen and Overeem 2013).

Finally, *there is no best method of assessment*; there is no magical bullet that can do it all. Every single assessment approach or method will always be a compromise on its quality characteristics (Van der Vleuten 1996). No single method may cover the entire competency pyramid, no single method may be perfectly reliable, and no single method may provide all encompassing feedback. Good assessment requires

a combination of different assessment methods. To assess the entire competency pyramid, we need multiple methods and we need to combine information across multiple assessment sources.

28.3 Perspective 2: Assessment for Learning

The first assessment perspective discussed important features of assessment of learning. When students have passed a test that is used to make summative decisions, for example, certification, they tend to ignore the feedback coming from it (Harrison et al. 2013, 2014). The practical implications are manifold, both practically and conceptually. From a practical perspective, it is important to align the instructional objectives with the way the assessment is done. When curriculum objectives and the assessment are misaligned, the assessment generally prevails, in many times leading to undesirable learning strategies by students. It is important therefore to verify the effect of assessment on the learning of students. From a conceptual viewpoint, the close link between assessment and learning invites us to think about assessment as part of the instructional arrangement and as part of the learning process (Boud and Falchikov 2007). This is the assessment *for* learning concept, the second assessment perspective in this chapter. Viewing assessment as an educational design problem has phenomenal consequences. How does the assessment fit in the instructional plan? How is feedback given and used? What does the scheduling need to be in time? How do methods map to certain competencies to be achieved? And so on. Assessment as an educational design issue needs further attention and more research. In the following, we will clarify the concept of assessment for learning and discuss strategies that teachers and students can use in making assessment beneficial and motivating for learning.

28.3.1 Strategies to Enhance Assessment for Learning

To define assessment for learning, we will refer to the definition of the Assessment Reform Group: 'Assessment for learning is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning' (Assessment Reform Group 2002). This definition implies that assessment for learning can vary on a continuum from informal 'on the fly' activities – which means that teachers and students continuously interact during the assessment process – to formal embedded activities that are consciously organized, for example, a self- or peer assessment (Shavelson et al. 2008). Particularly the following situations are illustrative for assessment *for* learning (Clark 2010):

– Students are engaged in a process that focuses on metacognitive strategies.

Table 28.2 Strategies to enhance formative assessment

Strategies	Role in formative assessment
Providing feedup, feedback and feedforward	Closing the gap between what learners already know and what they have to know by providing information to the learner that changes or stimulates behaviour
Rich questioning	Provides insight in learners' thinking to enable timely interventions, to refute misconceptions and promote deeper learning
Assessment dialogues	Effective for clarifying learning objectives and the establishment of criteria for success by scaffolding information. Helpful in gathering information about students' understanding and to ensure that students achieve the learning objectives
Reflective lessons	A well-considered combination of several assessments in one lesson to gather information about the development of the learners to choose the next step of instruction
Self-assessment	Provides the learner with information about his or her progression by relating products to learning objectives
Peer assessment	The involvement of peer learners in the assessment of a learner's progression stimulates the understanding of learning objectives and criteria
Rubrics	By describing the levels of attainment for different criteria, transparency is provided to communicate about criteria and expectations
Formative use of summative assessment	Evaluating summative assessments with students provides insight in what learners know and what they not yet know

– Students are supported in their efforts to think about their own thinking.
– Understand the relationship between their previous performance, their current understanding and clearly defined success criteria.
– Positioned as the agent improving and initiating their own learning.

Assessment for learning is often referred to as formative assessment. Formative assessments might be integrated in all learning situations and enhance all the teacher and learner activities that provide information that can be used to adjust learning. The most effective strategies for formative assessment are described in a thorough review (Sluijsmans et al. 2013). These strategies are summarized in Table 28.2. We will discuss four strategies in more detail: feedback, self-assessment, peer assessment and rubrics.

28.3.2 Feedback

Feedback is seen as the most effective method to make assessments formative. Formative feedback represents information communicated to the learner that is intended to modify the learner's thinking or behaviour for the purpose of improving learning (Shute 2008). Feedback becomes formative when students are provided

with scaffolded instruction or thoughtful questioning that served as a prompt for further enquiry, which then closes the gap between their current level of understanding and the desired learning goal (Clark 2010). Effective feedback includes two types of information: verification and elaboration. Verification is defined as the simple judgement of whether an answer is correct (also known as knowledge of results). Elaboration is the informational – and therefore more formative – aspect of the message that provides relevant cues to guide the learner towards a correct answer. Feedback will be effective if it answers three questions asked by a teacher and/or by a learner: Where am I going? How am I going? and Where to next? (Hattie and Timperley 2007). This gives subsequently information about the understanding of the learning objectives (feed up), information about the progress made towards the learning objectives (feedback) and information about the activities that need to be done to make progress (feed forward).

Providing feedback may not always be effective. There should be a coordinated plan with clear and decisive statements regarding feedback scope and functions, content, timing, presentation, conditions and the actors (Narciss 2013). Black and Wiliam distinguish three main actors playing a role in answering these three questions: the teacher, the learners and the peers (Black and Wiliam 2009). The teacher plays a role in clarifying and sharing learning intentions and criteria for success, engineering effective classroom discussions and other learning tasks that elicit evidence of learners understanding and providing feedback that moves learners forward. The peers are instructional resources for one another, and the learner is the owner of his own learning.

Shute (2008) provides a very interesting overview of guidelines that can be taken into account in feedback practice. This overview presents guidelines in terms of: things to do, things to avoid, timing issues and learner characteristics.

28.3.3 Self-Assessment

To become successful lifelong learning professionals, students are required to keep up with the latest developments in their expertise and to engage in a variety of tasks that foster continuous self-improvement (Bjork 1999; Boud 2000). Excellent sportsmen serve as a classic example of learners who continuously improve their performance by setting new goals and persisting to achieve these. Self-assessment is effective to foster self-improvement driven by the central question: How can I improve previous performances? (Eva and Regehr 2005). Self-assessment refers to the involvement of learners in making judgements about their own learning, particularly about their achievements and the outcomes of their learning (Boud and Falchikov 1989). Self-assessment is not a new technique, but a way of increasing the role of students as active participants in their own learning (Black and Wiliam 2009), and is mostly used for formative assessment in order to foster reflection on one's own learning processes and results. In a self-improvement model, students are presented with professional tasks that need to learn. For the self-assessment,

students are encouraged to select task aspects they would like to improve. After the assessment, students can self-assess their performance based on the selected task aspects and predefined standards. A new learning cycle starts when further development is needed. Most surprisingly, in research on self-assessment, this cyclic process of self-improvement is rarely used (Falchikov and Boud 1989; Gordon 1991). As a consequence, professional development becomes a unique learning path for each individual learner (Handfield-Jones et al. 2002). However, self-improvement assumes self-regulated learners, who are able to self-assess their performance (Boud 1990; Zimmerman 1990). Unfortunately, to date there is little empirical evidence that (starting) professionals are indeed capable of continually self-regulating their self-improvement process (Kruger and Dunning 1999; Regehr and Eva 2006). Abundant research has shown that we are poor self-assessors (Eva and Regehr 2005), so self-assessment should always be triangulated to other forms of assessment.

28.3.4 Peer Assessment

Falchikov defines peer assessment as the process through which groups of individuals rate their peers (Falchikov 1995). This exercise may or may not entail previous discussion or agreement over criteria. It may involve the use of rating instruments or checklists which have been designed by others before the peer assessment exercise or designed by the user group to meet its particular needs. Extensive literature reviews show that peer assessment is predominantly quantitative, such as peer ranking, peer nomination and peer rating (Sluijsmans et al. 2002; Van Zundert et al. 2010). However, for peer assessment to be effective for learning, it is recommendable that peer assessment is approached as a complex professional skill that requires intensive training. An example how students can be trained in peer assessment skills and how this affects learning can be found in Sluijsmans et al. (2002). First, they conducted a literature review and expert interviews to make an overview of the important peer assessment skills. This resulted in a peer assessment model in which three main skills are taken into account. These skills are (1) defining assessment criteria, thinking about what is required and referring to the product or process; (2) judging the performance of a peer, reflecting upon and identifying the strengths and weaknesses in a peer's product and writing an assessment report; and (3) providing feedback for future learning, giving constructive feedback about the product of a peer. Subsequently, a training programme for peer assessment was designed according to instructional design principles. This training programme consists of a number of peer assessment tasks, which are fully embedded in the curriculum. The study revealed that the qualitative assessment reports written by the students showed that the experimental groups surpassed the control groups in the quality of the assessment skill. As a result of the training, students from the experimental groups also scored significantly higher grades for the end products of the course than students from the control groups.

28.3.5 Rubrics

For constructive alignment, rubrics are a good tool to support assessment *for* learning as well as assessment *of* learning. Rubrics are very suited to assessment for learning as they can be made or adapted for many levels and provide both criteria and learning objectives. Although they appear in many forms and sizes, the best-known type is a grid, which allows teachers to align levels of performance with criteria by using descriptors. The descriptive language that explains characteristics of work or performance at increasing levels of quality makes rubrics informative tools for feedup, feedback and feed forward. By using them in a general way, they become efficient tools in assessment programmes as they judge quality across similar tasks for different courses. Rubrics allow teachers to communicate about standards and aims in a coherent and clear way. The transparency provided by rubrics can be used to support feedback and self- and peer assessment. Providing a rubric with samples of strong and weak work increases the transparency. A rubric provides insight in the complexity of a task and helps learners in answering the questions of where they are and what they have to do next to achieve higher quality (Burke 2008). In order to develop a feeling of ownership, it is a good idea to try and develop a rubric together with teachers and learners. Bottom-up use of rubrics, which implies the development of a rubric together with teachers and learners, is preferred above top-down use of rubrics. The bottom-up use results in feelings of ownership and a more and better understanding of the learning objectives (Burghout 2012). Arter and Chappuis identify the following basic steps in developing rubrics: (1) identify aims/learning objectives, (2) identify observable attributes that you want to see (and don't want to see) demonstrated, (3) brainstorm characteristics that describe each attribute and find ways to describe levels for each attribute, (4) write precise descriptors for lowest level and highest level, (5) write descriptors for the remaining (intermediate) levels and (6) collect samples of work which exemplify each level to become benchmarks (Arter and Chappuis 2006). Providing feedup, feedback and feed forward in combination with the use of rubrics is a strong combination of tools for assessment for learning. It can be used in supporting other formative methods like self-assessment, peer assessment, reflective lessons and assessment dialogues.

28.4 Perspective 3: Assessment as Learning

28.4.1 A Programmatic Approach Where Assessment of and for Learning Are Merged

The literature on assessment is virtually all geared towards the study of individual assessment methods and how to optimize them. As is clear from the above, every individual method of assessment has clear limitations and is far from perfect. This led to the suggestion to not optimize the individual method but to optimize the

collection of methods in a programme of assessment (Van der Vleuten and Schuwirth 2005). In a programme of assessment, methods of assessment are purposefully selected, mainly because of their intended positive effect on learning (Schuwirth and Van der Vleuten 2011). A curriculum is a good metaphor. In the past a curriculum consisted of individual teachers or departments each doing their course or module with little integration across courses. The set of courses made up the curriculum, and the integration was left to the learner when confronted to professional tasks or to work. Modern curricula are planned according to a master plan, are integrated, are governed, are evaluated and are modified accordingly. The same strategy can be taken with assessment. Within a curriculum a set of assessment methods are chosen deliberately and coherently based on intended learning effects. Any method of assessment may be used (classic or modern) all depending on its purpose within the total assessment programme. For example, one deliberately requires learners to present here, to verbalize there, to synthesize then, to write up subsequently and so on so forth. The programme is evaluated regularly and modified accordingly. After the suggestion to start thinking in assessment programmes, Baartman et al. developed a set of criteria of assessment programmes (Baartman et al. 2006) and a self-evaluation instrument to judge the quality of these programmes (Baartman et al. 2011). Dijkstra et al. (2012) continued to develop a set of 73 guidelines for designing assessment programmes (Dijkstra et al. 2012). Van der Vleuten et al. (2012) proposed a model of what has been called programmatic assessment. We will describe this model and end with an illustration of an implementation of the model in a concrete setting.

28.4.2 Key Elements of Programmatic Assessment

In programmatic assessment, a number of basic tenets are formulated that are inspired on the set of principles that are formulated earlier (Table 28.1):

- Each individual assessment moment is but one datapoint.
- Each individual datapoint may consist of any method, is closely linked to the educational programme and always provides meaningful feedback to the learner.
- A continuum of stakes replaces the formative-summative distinction.
- Stakes and number of datapoints are proportionally related; a single datapoint is by definition low stake.
- Information is aggregated across datapoints and across meaningful entities (usually a competency framework).
- Learners are supported in feedback use and follow-up (usually through a mentoring system).
- High-stake (promotion or selection) decisions are based on many datapoints and are taken by a (independent) committee.

In this model individual datapoints of assessment are maximally optimized for learning. Every assessment provides feedback to the learner in whatever way that is

appropriate within that method (quantitative and/or qualitative). The individual datapoint is low stake. Usually, pass/fail decisions are not taken. The individual datapoint serves to provide information on the learner, not on pass/fail decision-making. The decision-making on passing or failing is done at the programme level when many datapoints are gathered. High-stake decisions, for example, promotion to the next year, are ideally preceded by intermediate decisions on learner progress. The high-stake decision-making should not really come as a surprise to the learner. Learners reflect on their feedback and progress. They are supported in that process through a dialogue, usually a mentor. Progress is discussed regularly and plans are made for further study. This may include remediation on certain elements that have shown insufficient progress. The collection of datapoints is meaningfully aggregated, for example, across certain competencies and assessed by a committee of experts. Committee deliberation will be proportional to the clarity of the information being assessed. If the assessment information triangulates in a clear picture (which will be the case for the far majority of learners), then the decision-making will be a swift process. On the other hand, more difficult cases will require the committee to deliberate extensively. The ultimate decision taken can be justified and defended by the committee.

28.4.3 A Best Practice of Programmatic Assessment

An illustration may provide more clarity on programmatic assessment. Maastricht University has a graduate entry programme in medicine. This is a 4-year training programme in medicine and research. On top of the medical degree, students receive a Master of Science degree in clinical research. Matriculating students have a previous bachelor or master degree in any of the biological sciences (n = 50). The first 2 years consist of theoretical training through problem-based learning, the latter 2 years of clinical work-based learning and research projects. The curriculum and the assessment are structured according to a competency framework (CanMEDS) (Frank and Danoff 2007). Within units, the assessment consists of traditional end-of-unit assessment (in written or oral form), assignments and projects and peer and tutor assessment. All assessment is feedback oriented. There are no pass/fail decisions. On top of the modular assessment, there is continuous longitudinal assessment. One form is through progress testing (Wrigley et al. 2012). A progress test is a comprehensive multiple-choice test representing the end objectives of medical training and contains all disciplines. Most items are problem-oriented scenarios. One could compare a progress test to a final examination, but that final examination is given to all the students in the curriculum. This is repeated four times per year. Every new test has newly written items. Students cannot really prepare for a progress test. What would they study? Anything can be asked. There is also no need to study for it as well. If a student regularly studies, scores will grow automatically. Every 3 months all students can see how they have grown in total and in all parts of the blueprint (disciplines and organ system categories). Feedback is provided

through an online system where students may analyse their (longitudinal) performance in any area. Progress testing provides a wealth of information to the learner on their achievements in the cognitive domain, while at the same time prevents test-directed studying for the short term. Another longitudinal assessment is the assessment on their competencies through self, peer and tutor assessment. During their workplace attachments, they are regularly observed and feedback is given. A number of different work-based instruments are used (direct observation instruments, multisource feedback, video assessment, assignments and project assessments). The student assembles all the feedback in an electronic portfolio. The portfolio is a portal that serves as a repository, an organizer for feedback reception (e.g. forms may be completed on hand-held devices; feedback questionnaires are distributed to assessors), and aggregates information across different assessment sources into competency-based quantitative and qualitative feedback reports. Every student has a mentor who has access to the portfolio. The mentor and the student regularly meet. Progress is discussed and (remediation) plans are being made and monitored. At the end of the academic year, the portfolio is assessed for promotion to the next year. The mentor writes a recommendation that may be annotated by the student. The pass/fail (and distinction) decision is given by a committee. The committee consists of all mentors, but in the actual decision-making, the own mentor of the student has no say. The committee extensively deliberates only on a few cases. Students may appeal to committee decisions.

The programmatic assessment model optimizes both learning and decision-making. Learning is optimized because the assessment is information rich. Measures are taken that facilitate the use of feedback. Self-directed learning is facilitated by continuous reflection and feedback. Remediation, as opposed to the classic repetition (in resits or in redoing the course), is an on-going and personalized process. The decision function is removed from the individual datapoint, and therefore reliability is of less concern. The biggest concern at the individual datapoint is the provision of meaningful information for learning. The decision-making function is optimized due to the use of many (rich) datapoints. Such richness of information can never be replaced by any other single assessment such as a final examination. Programmatic assessment has been implemented in a number of settings across the world, albeit in health-related fields so far (Dannefer 2013; Bok et al. 2013). The first research shows that programmatic assessment may work in higher education (Bok et al. 2013; Heeneman et al. Under editorial review), but the quality of implementation is kernel to its success. Getting high-quality feedback from teachers or fieldwork supervisors is a challenge. Creating and getting buy-in from teachers are important. The users (both teachers and learners) should understand the assessment function and their role in the assessment. Convincing teachers, for example, that they may not fail students is not an easy task. Cost is another issue to consider. Giving feedback takes time; a mentorship programme is resource intensive. Programmatic assessment should therefore be part of a full curriculum concept, where the assessment is part of the learning concept. When fully integrated in the learning concept, assessment as learning will be achieved. When implemented carefully, learners tend to become real feedback seekers (Altahawi et al. 2012).

28.5 Conclusions

In this chapter we outlined three perspectives on assessment and their implications for educational practice: assessment *of* learning – with a focus on making sound and reliable decisions about students' learning – and assessment *for* and *as* learning, both focused on enhancing students' further professional learning. Figure 28.2, which is based on the model presented by Clark (2010), summarizes the main messages regarding each perspective in relation to the triangle assessment, curriculum and learning/teaching. From Fig. 28.2, two important conclusions can be drawn. A first conclusion is that assessment of, for and as learning can and should be aligned within the whole educational system. When this alignment is assured, appropriate decisions about students' professional learning can be made which are also informative for subsequent learning in which students are seen as active participants who are made responsible for their own learning. A second conclusion is that strong assessment practice involves intense collaboration between students, staff and

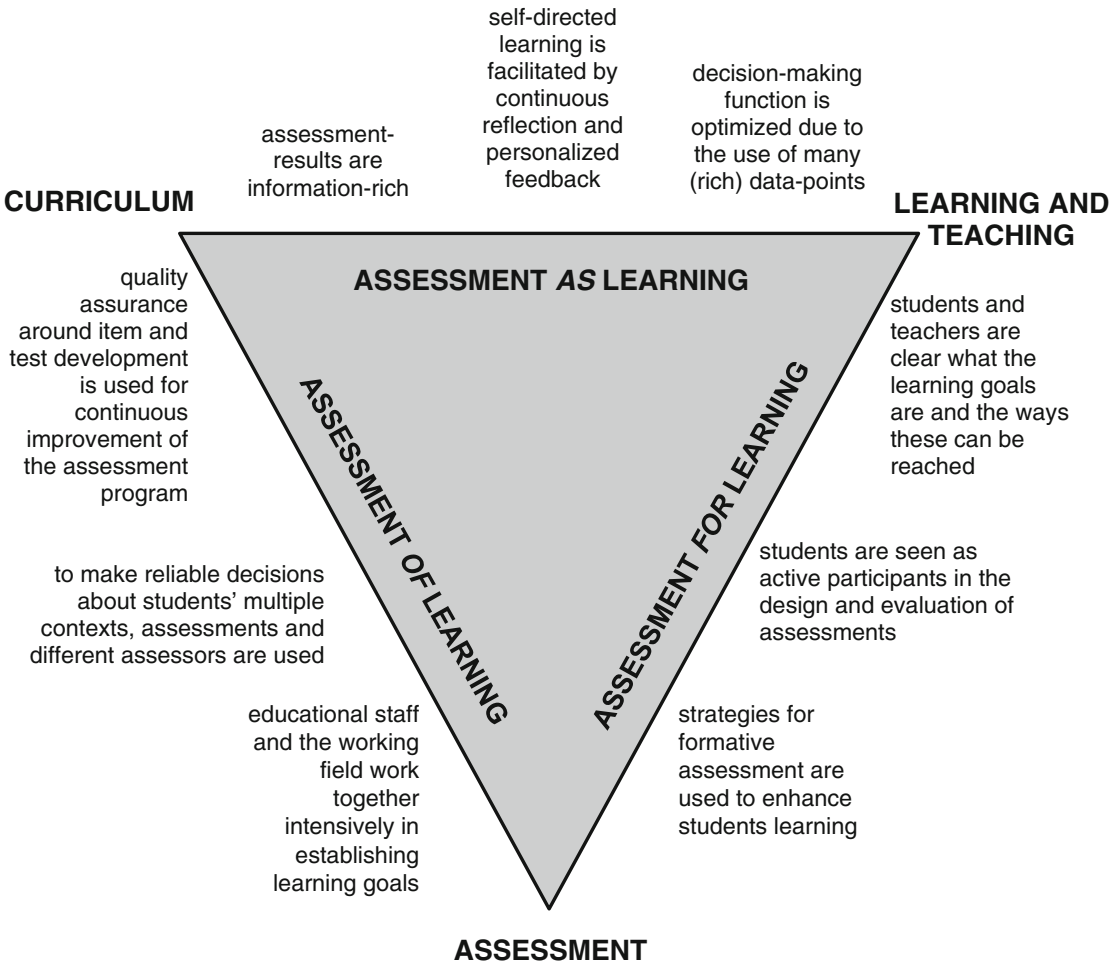

Fig. 28.2 Assessment of, for and as learning to assure constructive alignment (Clark 2010; Adapted by the authors)

working field. Since many professions evolve rapidly, a well thought-out quality assurance system is needed in which learning goals, and the consequences for assessment design are evaluated on a regular basis.

We hope we clarified our basic view on how assessment may support the learner and the learning process. In all sectors of education and in all parts of the world, competency-based assessment is being introduced. Our educational philosophy has moved from a behaviourist view on learning to a constructivist view on learning. But many assessment practices are still behaviouristic and more appropriate for a mastery learning conception. Given the driving effect of assessment on learning, we would argue that the success of introducing competency-based education depends on the quality of the assessment in such education. Competency-based education addresses learning of complex skills. Competencies typically have a behavioural aspect that can only be assessed through observation and professional judgement. Such skills can only be developed with vertical integration in longitudinal lines of training with proper feedback cycles and follow-up. Assessment may provide and should provide the right scaffold for this learner support. To be successful, assessment should be part of the design process of a training programme, intrinsically linked to the on-going education, providing feedback for learner support and providing accumulating information for monitoring and decision-making over learner progress.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876.

Altahawi, F., Sisk, B., Poloskey, S., Hicks, C., & Dannefer, E. F. (2012). Student perspectives on assessment: Experience in a competency-based portfolio system. *Medical Teacher, 34*(3), 221–225. doi:10.3109/0142159X.2012.652243.

Arter, J., & Chappius, J. (2006). *Creating & recognizing quality rubrics*. Portland: Educational Testing Service (ETS).

Assessment-Reform-Group. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.

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 programmes. *Studies in Educational Evaluations, 32*(2), 153–170.

Baartman, L. K., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. (2011). Self-evaluation of assessment programs: A cross-case analysis. *Evaluation and Program Planning, 34*(3), 206–216. doi:10.1016/j.evalprogplan.2011.03.001.

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances in Health Sciences Education, 18*(4), 559–571.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 453–459). Cambridge, UK: MIT Press.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5–31.

Bok, H. G., Teunissen, P. W., Favier, R. P., Rietbroek, N. J., Theyse, L. F., Brommer, H., et al. (2013). Programmatic assessment of competency-based workplace learning: When theory meets practice. *BMC Medical Education, 13*(1), 123.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*(1), 101–111.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151–167.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*(5), 529–549.

Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. London: Routledge.

Brett, J. F., & Atwater, L. E. (2001). 360° feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology, 86*(5), 930.

Burghout, C. (2012). We do need some education in assessment for learning. In *Keynote talk for the TED schools conference on linking assessment to learning*. Turkey: Istanbul.

Burke, K. (2008). *How to assess authentic learning* (5th ed.). Thousand Oaks: Corwin.

Butler, P. (2007). *A review of the literature on portfolios and electronic portfolios*.

Chappell, C., Gonczi, A., & Hager, P. (2000). *Competency-based education*.

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. J., & van der Vleuten, C. P. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. [Research Support, Non-U.S. Gov't]. *Advances in Health Sciences Education: Theory and Practice, 17*(1), 39–53. doi:10.1007/s10459-011-9292-5.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah: Lawrence Erlbaum.

Clark, I. (2010). Formative assessment: 'There is nothing so practical as a good theory'. *Australian Journal of Education, 54*(3), 341–352.

Dannefer, E. F. (2013). Beyond assessment of learning toward assessment for learning: Educating tomorrow's physicians. *Medical Teacher, 35*(7), 560–563.

Dijkstra, J., Galbraith, R., Hodges, B. D., McAvoy, P. A., McCrorie, P., Southgate, L. J., et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*, 20. doi:10.1186/1472-6920-12-20.

Dochy, F., Segers, M., Gijbels, D., & Struyven, K. (2007). Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 87–100). Oxford: Routledge.

Driessen, E. W., & Overeem, K. (2013). Mentoring. In K. Walsh (Ed.), *Oxford textbook of medical education* (pp. 265–284). Oxford: Oxford University Press.

Driessen, E. W., Van der Vleuten, C. P. M, Schuwirth, L. W. T., Van Tartwijk, J., & Vermunt, J. D. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39*, 214–220.

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine, 80*(10 Suppl), S84–87. doi:80/10_suppl/S84 [pii].

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353–371.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*(10 Suppl), S46–S54.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Programmed Learning, 32*(2), 175–187.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.

Feltham, R. (1988). Validity of a police assessment centre: A 1–19-year follow-up. *Journal of Occupational Psychology, 61*(2), 129–144.

Fokkema, J. P., Teunissen, P. W., Westerman, M., van der Lee, N., van der Vleuten, C. P., Scherpbier, A. J., et al. (2013). Exploration of perceived effects of innovations in postgraduate medical education. *Medical Education, 47*(3), 271–281.

Frank, J. R., & Danoff, D. (2007). The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. *Medical Teacher, 29*(7), 642–647. doi:787705294 [pii]10.1080/01421590701746983.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*(12), 762–769.

Govaerts, M. J. B., & Van der Vleuten, C. P. M. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164–1174.

Govaerts, M. J., van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2005). The use of observational diaries in in-training evaluation: Student perceptions. [Evaluation Studies]. *Advances in Health Sciences Education, 10*(3), 171–188. doi:10.1007/s10459-005-0398-5.

Govaerts, M. J., Van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239–260.

Handfield-Jones, R. S., Mann, K. V., Challis, M. E., Hobma, S. O., Klass, D. J., McManus, I. C., et al. (2002). Linking assessment to learning: A new route to quality assurance in medical practice. *Medical Education, 36*(10), 949–958.

Harrison, C. J., Könings, K. D., Molyneux, A., Schuwirth, L. W., Wass, V., & van der Vleuten, C. P. (2013). Web-based feedback after summative assessment: How do students engage? *Medical Education, 47*(7), 734–744.

Harrison, C. J., Könings, K. D., Schuwirth, L., Wass, V., & van der Vleuten, C. (2014). Barriers to the uptake and use of feedback in the context of summative assessment. *Advances in Health Sciences Education, 48*, 1–17.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Heeneman, S., Oudkerk-Pool, A., Schuwirth, L., van der Vleuten, C., & Driessen, E. (Under editorial review). *Students as active constructors of their learning while assessing at the program level: The concept versus practice*.

Heijke, H., & Meng, C. (2006). *Discipline-specific and academic competencies of the higher educated: their value in the labour market and their acquisition in education* (No. ROA-W-2006/9E). Maastricht: Research Centre for Education and the Labour Market.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment centre*. *Journal of Occupational Psychology, 64*(1), 1–21.

Jozefowicz, R. F., Koeppen, B. M., Case, S. M., Galbraith, R., Swanson, D. B., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156–161.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. doi:319/5865/966 [pii]10.1126/science.1152408.

Koutselini Ioannidou, M. (1997). Testing and life-long learning: Open-book and closed-book examination in a university course. *Studies in Educational Evaluation, 23*(2), 131–139.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43*(1), 21–27.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513.

Merrienboer, J., & Kirschner, P. (2007). *Ten steps to complex learning. A systematic approach to four-component instructional design*. New York/London: Routledge.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67.

Moonen-van Loon, J., Overeem, K., Donkers, H., van der Vleuten, C., & Driessen, E. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087–1102.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal. Social, organizational and goal-based perspectives*. Thousand Oaks: Sage.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review, 23*, 7–26.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher, 29*(9), 855–871.

Pelgrim, E. A., Kramer, A. W., Mokkink, H. G., van den Elsen, L., Grol, R. P., & van der Vleuten, C. P. (2011). In-training assessment using direct observation of single-patient encounters: A literature review. [Review]. *Advances in Health Sciences Education: Theory and Practice, 16*(1), 131–142. doi:10.1007/s10459-010-9235-6.

Petrusa, E. R. (2002). Clinical performance assessments. In G. R. Norman, C. P. M. Van der Vleuten, & D. I. Newble (Eds.), *International handbook for research in medical education* (pp. 673–709). Dordrecht: Kluwer Academic Publisher.

Regehr, G., & Eva, K. (2006). Self-assessment, self-direction, and the self-regulating professional. *Clinical Orthopaedics and Related Research, 449*, 34–38.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*(1), 103–108.

Sargeant, J., Mann, K., Sinclair, D., Van der Vleuten, C., & Metsemakers, J. (2008). Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education, 13*(3), 275–288.

Sargeant, J. M., Mann, K. V., van der Vleuten, C. P., & Metsemakers, J. F. (2009). Reflection: A link between receiving and using assessment feedback. [Research Support, Non-U.S. Gov't]. *Advances in Health Sciences Education: Theory and Practice, 14*(3), 399–410. doi:10.1007/s10459-008-9124-4.

Schuwirth, L. W., & van der Vleuten, C. P. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974–979. doi:10.1111/j.1365-2929.2004.01916.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*(6), 478–485.

Segers, M., & Dochy, F. (2001). New assessment forms in a problem-based learning: The value-added of the students' perspective. *Studies in Higher Education, 26*, 327–343.

Segers, M., Dochy, F., & Cascallar, E. (Eds.). (2003). *Optimising new modes of assessment: In search for qualities and standards*. Dordrecht: Springer.

Semeijn, J. H., van der Velden, R., Heijke, H., van der Vleuten, C., & Boshuizen, H. P. (2006). Competence indicators in academic education and early labour market success of graduates in health sciences. *Journal of Education and Work, 19*(4), 383–413.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., et al. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295–314.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Sluijsmans, D. M. A., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454.

Sluijsmans, D. M. A., Joosten-ten Brinke, D., & Van der Vleuten, C. P. M. (2013). *Toetsen met leerwaarde. Een reviewstudie naar effectieve kenmerken van formatief toetsen* [Assessment for learning: What are methods and conditions of formative assessment that are effective for learning?]. The Hague: NWO.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*(5), 5–11.

Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education, 1*(1), 41–67.

Van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317. doi:10.1111/j.1365-2929.2005.02094.x.

Van der Vleuten, C. P., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. [Review]. *Medical Education, 25*(2), 110–118.

Van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. [Review]. *Best Practice & Research. Clinical Obstetrics & Gynaecology, 24*(6), 703–719. doi:10.1016/j.bpobgyn.2010.04.001.

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

Van Tartwijk, J., & Driessen, E. W. (2009). Portfolios for assessment and learning: AMEE Guide no. 45. *Medical Teacher, 31*(9), 790–801.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279.

Verhoeven, B., Verwijnen, G., Scherpbier, A., Schuwirth, L., & Van der Vleuten, C. (1999). Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education and Health, 12*(1), 49–60.

Watling, C., Driessen, E., Vleuten, C. P., Vanstone, M., & Lingard, L. (2013). Beyond individualism: Professional culture and its influence on feedback. *Medical Education, 47*(6), 585–594.

Wilkinson, J. R., Crossley, J. G., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education, 42*(4), 364–373.

Wood, L., Hassell, A., Whitehouse, A., Bullock, A., & Wall, D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design★. *Medical Teacher, 28*(7), e185–e191.

Wrigley, W., Van der Vleuten, C. P., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher, 34*(9), 683–697. doi:10.3109/0142159X.2012.704437.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17.

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 647–679). Westport: American Council on Education/Praeger.