Chapter 45
Assessment of Professional Competence

Dineke E.H. Tigelaar and Cees P.M. van der Vleuten

Abstract This chapter deals with research on judging, evaluating, monitoring and assessing professional competence in educational contexts. In line with current views on assessment and learning, we argue that assessment can and should be used to develop professional competence. Drawing on research from different areas of professional competence, we extend this line of argumentation by contending that learning and instruction can benefit when different methods of assessment are strategically combined in a coherent assessment programme. We will discuss the optimization of education- and practice-related purposes, formative and summative purposes and quality issues in the assessment of professional competence, and conclude with some prospects for supporting expert judgement, developing guidelines for assessment programmes and gaining improved understanding of mechanisms underlying the impact of assessment on learning.

Keywords Professional competence • Assessment methods • Assessment programs

D.E.H. Tigelaar (✉)
Graduate School of Teaching, ICLON-Leiden University,
P.O. Box 905, 2300 AX Leiden, The Netherlands
e-mail: dtigelaar@iclon.leidenuniv.nl

C.P.M. van der Vleuten
Department of Educational Development and Research, School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University,
PO Box 616, 6200 MD Maastricht, The Netherlands
e-mail: c.vandervleuten@maastrichtuniversity.nl

S. Billett et al. (eds.), *International Handbook of Research in Professional and Practice-based Learning*, Springer International Handbooks of Education,
DOI 10.1007/978-94-017-8902-8_45,

45.1 Introduction

We discuss from a research perspective, the assessment of professional competence in educational settings. In this chapter we focus on psychological theories in addressing the assessment of professional competence. Professional competence is a complex concept that allows for various definitions. In their seminal paper on the concept of competence, Stoof et al. (2002) stressed that definitions of competence are purpose and context dependent, although this does not preclude commonalities between contexts. One characteristic of professional competence that is shared across domains is the ability of professionals to fulfil complex core tasks by integrating and applying appropriate context- and domain-specific knowledge, skills and attitudes (Stoof et al. 2002). Developers of methods to assess professional competence in educational contexts should define not only professional functioning for the context of interest but also ways of promoting its development. We agree with Stoof et al. (2002) that definitions of competence are purpose- and context-dependent, but we also acknowledge that definitions of professional competence for different areas share important characteristics, such as relevance to the domain of possible encounters within a particular area of practice which a professional is expected to manage effectively (Kane 1992) and the knowledge, skills and judgement the professional is expected to use in managing these encounters.

In the literature several purposes that can be served by assessment of professional competence have been identified. Firstly, assessment can be a selection tool for admission to education and training programmes or job appointments. Secondly, assessment of professional competence can be used to monitor and guide the progress of learners through a training programme. Thirdly, it can be used to determine whether students meet requirements for licensure to practise in their chosen profession and whether staff members qualify for promotion, and finally, results of assessment can be used to monitor the effectiveness of education and training programmes and provide evidence for programme quality. These four purposes: selection, diagnosis, licensure and accountability are usually subdivided into (a) formative assessment focused on diagnosing and monitoring student performance in order to enhance learning and development and (b) summative assessment focused on decision making for selection, licensure and accountability.

We adhere to current views on assessment and learning that assessment for educational purposes can and should concomitantly serve to stimulate learners to develop their professional competence. This viewpoint is consistent with the notion that assessment drives learning (Frederiksen 1984), as reflected in the more recent idea of the integration of assessment, instruction and learning. Integration requires substantial consistency of learning, instruction and assessment, often referred to as the principle of ‘alignment’ (Biggs 1996). Taking our line of reasoning one step further, we argue that assessment should be used strategically within a programme of assessment so as to maximize its potential as a tool for learning and instruction.

We aim to demonstrate that effective strategic use of assessment in educational contexts relies on carefully chosen methods of assessment combined in an

assessment programme (Chester 2003; Van der Vleuten and Schuwirth 2005). In line with Van der Vleuten and Schuwirth (2005) and Baartman et al. (2007), we view assessment as an *instructional design problem*, covering the full range of assessment methods used within a curriculum. An effective programme for assessing professional competence should be informed by professional practice in the domain of interest. Since professional practices differ across professions and settings, assessment of professional competence cannot be treated as an entity that is uniform across all educational contexts. Given this inherent variability, programmes for the assessment of professional competence depend crucially on well-informed choices made by programme designers based on their knowledge both of the profession and of assessment. Despite inter-contextual differences, developers of assessment programmes for professional competence can learn a great deal from studying existing assessment programmes. We therefore present a case from teacher training as a thick description which may be helpful to readers in their deliberations about the implementation of a programmatic approach to assessment in their own context.

The ideas presented in this chapter draw on overviews of assessment of professional competence, published mainly in the context of medical education (Van der Vleuten 1996; Van der Vleuten and Schuwirth 2005; Schuwirth and Van der Vleuten 2011; Van der Vleuten et al. 2012) but include also insights from other areas of professional practice, teaching in particular.

This chapter is written for (a) educators who are interested in assessment of professional competence, especially in relation to the development and improvement of assessment systems; and (b) administrators, supervisors and educators involved in the development of programmes for the assessment of professional competence in their own curricula.

45.2 What Is Professional Competence? Implications for Assessment

Definitions of professional competence rely on the specific requirements of the profession in question and the knowledge, skills and judgement to be mastered by professionals to be able to manage professional encounters in their field. Relevant notions regarding the assessment of professional competence will be presented in a simple and well-known model from medical education.

45.2.1 What Is Professional Competence?

Competence has long been conceptualized – implicitly – as comprising several distinct components (Van der Vleuten 1996), each to be mastered separately in monotonic process driven by learning experiences. These components were

considered to be relatively stable across situations and time, reflecting the *trait conception* which was generally supported in psychology at that time. According to the trait conception, relatively stable characteristics cause individuals to behave in certain ways (Van der Vleuten 1996), and components of competence are related to a set of latent factors within an individual, which affect performance but cannot be observed and therefore have to be inferred from observed behaviours. Between 1950 and 1980, as a result of the cognitive revolution in psychology (Baars 1986), the focus of research shifted from observable behaviours to the unobservable workings of the professional's mind. Studies investigated thought processes, decision-making and planning, conceptualizations and problem solving. Concepts of competence now also encompassed cognitive aspects, such as how professionals apply their knowledge in a particular context. It was acknowledged that professionals often have to tackle problems for which no straightforward solution is available but which require novel combinations of skills and knowledge (Andrews and Barns 1990). To deal with ill-defined problems, professionals draw on a unique knowledge base and a situational life space which enable them to exercise decision making skills (Andrews and Barns 1990; Darling-Hammond et al. 1983). In line with these insights, Kane (1992) proposed that conceptions of competence involve two major components: the domain of possible encounters within a particular area of practice which a professional needs to manage effectively, and the knowledge, skills and judgement that are prerequisite for managing those encounters. Kane's conceptualization has been embraced by various authors, including Roelofs and Sanders (2007), who defined competent teaching as:

> the extent to which a teacher, as a professional, takes deliberate and appropriate decisions (based on personal knowledge, skills, conceptions, etc.), within a specific and complex professional context (students, subject matter, etc.), resulting in actions which contribute to desirable outcomes, all according to accepted professional standards. (in Bakker et al. 2011, p. 125).

According to this definition, assessment of teaching competence covers teachers' decision-making processes and actions and how these impact on student learning. Note that this definition combines cognitive reasoning by referring to deliberate decisions and actions, as well as situational awareness, by focusing on appropriate decisions and actions. In the introduction we explained that professional competence comprises general characteristics as well as specific characteristics of professional contexts, such as business organizations and educational institutions. So definitions of professional competence share commonalities and at the same time differ across professional contexts. Using a simple model of professional competence for illustration, we will discuss insights from research regarding the nature of professional competence and suitable assessment modalities.

45.2.1.1 A Simple Model for Assessing Professional Competence

The above-mentioned simple model of competence assessment was introduced by Miller in 1990 in the context of medical education. The model is visualized as a pyramid, which from the ground level upwards consists of four consecutive layers

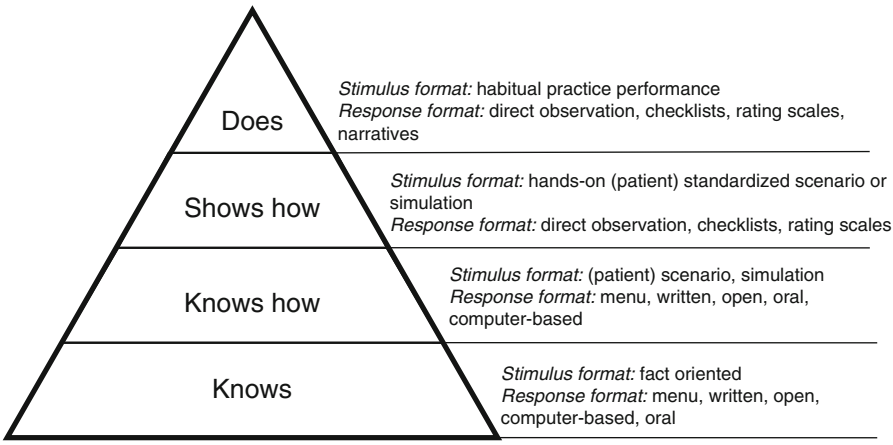


Fig. 45.1 Miller's pyramid explained

denoting increasing competence (Fig. 45.1), with types of assessment that are particularly suited to specific layers. The bottom layer deals with factual knowledge and the next layer with knowing how to apply that knowledge, labelled as the 'knows' and the 'knows how' level, respectively. Performing all aspects of the competence of the 'knows how' level in a controlled simulated or laboratory environment comprises the third layer, the 'shows how' level. The top layer is concerned with the 'does' level, i.e. performance in an authentic setting in day-to-day professional practice. Miller did not design his pyramid as a pedagogical model where mastering the highest level is conditional on mastery of the lower levels, but the pyramid provides anchor points for decisions about the suitability of assessment methods for different contexts. The cognitive skills for the 'knows' level are mostly assessed by written, oral or computer-assisted knowledge tests, while the upper levels require concrete evidence of mastery of psychomotor, cognitive and affective components of competence, usually assessed using direct observation of performance or evaluation of products of performance.

We will refer to Miller's pyramid in describing different methods and approaches for the assessment of professional competence. We will also show how research on the quality of these assessment methods and approaches has yielded new understandings of the nature of professional competence, which have inspired new assessment technologies and theories.

45.3 A Historical Overview of Developments in the Assessment of Professional Competence

We present an overview of developments and research related to the assessment of professional competence, ending with a summary of implications and principles of assessment, illustrating how assessment of professional competence has developed

into an enterprise involving the use of multiple methods to support feedback to guide self-directed learning as well as methods to arrive at sound judgements to support defensible and trustworthy decisions about certification and promotion.

45.3.1 Developments in Assessment Methods

The '*knows*' level of Miller's pyramid is related to traditional paper and pencil factual knowledge tests and tests using multiple choice questions (Van der Vleuten 1996). Gradually the realization has dawned that knowledge may equally well be assessed using open-ended questions, oral examinations and computerized tests.

The '*knows how*' level requires evaluation of processes in people's minds, such as reasoning and problem solving (i.e. procedural knowledge). It is consequently important to make a distinction between the assessment task, i.e. the *stimulus format*, and the way candidates' responses are captured, i.e. the *response format*. A stimulus may be a written task eliciting a fact or a written case of a teaching situation inviting a candidate teacher to explain how they would deal with it, etc. Reponses may be captured by a write-in, a long text (essay), an oral situation, a menu of options (multiple-choice questions), etc. An important characteristic of assessment at the '*knows how*' level is that stimuli present a rich context, preferably derived from the professional domain. Such enriched tasks can be written cases, such as patient problems for medical students (McGuire and Babnott 1967) and they are intended to encourage candidates to consider the meaning of their knowledge and its application in concrete situations. The candidate's answers and decisions serve as indicators of the candidate's problem-solving ability. Computer techniques can be used to add realism to the assessment, e.g. by adding authentic videos and sounds.

On the '*shows how*' level, evaluation techniques are aimed at simulating real-life situations in the work or educational context. A knowledgeable assessor observes the candidate's actions and behaviours, often assisted by descriptions of adequate performance in the domain, based on definitions of key aspects of professional functioning and indicating what is considered to be good performance (Sadler 1987, 1989). Standards are also used to define levels of professional competence, from 'poor' to excellent'. A later development was *live* simulations, such as business simulation exercises in business education (Anderson and Lawton 1988), simulated clinical stations in medical education (Harden and Gleeson 1979) and assessment centres in teacher education (Shulman et al. 1988; Haertel 1991), with candidates completing a coherent set of tasks and activities representing key aspects of professional work. Simulation tasks for a candidate teacher may involve planning a lesson or evaluating student work. In live simulations, e.g. in assessment centres for student teachers, the level of task performance and the candidates' rationales for their actions are used as the basis for judging their knowledge and skills (Shulman et al. 1988).

Miller's '*does*' level calls for assessment of how professionals perform their daily tasks in realistic settings. Influenced by attempts to optimize the practical relevance assessment, which started in the 1990s, assessment developers have

sought to move beyond simulations and assess the '*does*' level in the real working environment. Such workplace or practice-based assessment is suitable for work placements and internships to monitor and evaluate student learning. The stimulus is the authentic context, whether work or school based, which, by definition, cannot be controlled (van der Vleuten et al. 2010). In work-based contexts, direct assessment prevails, while in school-based contexts indirect assessment methods tend to dominate. Direct assessment targets behaviour, observed either directly or in retrospect based on a candidate's previous interactions. Common methods are direct observation, followed by oral feedback from peers, colleagues and others, such as pupils in teacher evaluation, clients in business education and training, and case discussions of videotaped patient encounters or evaluation of performance by patients in medical practice. In the management literature such assessments are referred to as performance appraisals. The dominant response format is an observation structure, such as a global rating scale often complemented with additional space for assessors' narrative comments. Such observation structures are often implemented electronically, with communication by email or smart phone facilitating assessment and offering an attractive feedback format. 360-degree or multisource feedback from different stakeholders within candidates' work environment, sometimes combined with self-assessment, is a familiar example of this type of assessment. In educational settings, self-assessment and peer feedback are widespread. Indirect assessment at the '*does*' level is often seen in schools or educational institutions, and may include products or artefacts resulting from activities undertaken by the learner, with evidence obtained from multiple sources over time, such as information and feedback from different sources compiled in logbooks or portfolios. Portfolios are collections of evidence for professional competence, and their popularity in higher, continuing, professional and basic education has been on the increase since the early 1990s. The first portfolio for teacher assessment was introduced in the Teacher Assessment Project at Stanford University to support assessment of teacher competence with other information besides assessment-centre grades (Shulman et al. 1988). Influenced by the movement towards more authentic and meaningful assessment, the Stanford project assessed teacher performance in different practical contexts collecting the resulting information in a teaching portfolio. The notion of a portfolio was borrowed from architects' and artists' portfolios, i.e. real files containing samples of designs, drawings and paintings to present to potential clients (Bird 1990). Analogous to these portfolios, a teaching portfolio contains samples of a teacher's work collected over time across contexts (Wolf and Dietz 1998). As the information is longitudinal, the portfolio can be used to aggregate numerous samples of a candidate's performance collected over an extended period of time. Initially used mainly summatively, portfolios are currently also used formatively, which has affected their structure and content (Van Tartwijk et al. 2007). While portfolios originally contained evidential materials with, at best, some notes indicating what the material was, where it was collected, and why it was included in the portfolio, more extensive uses for portfolios have been developed, such as stimulating reflection on personal development and planning (Mansvelder-Longayroux et al. 2007).

45.3.2 Research Findings on the Assessment of Professional Competence

We present findings from research into the quality of assessment of various aspects of professional competence, showing how this type of research has influenced developments in this type of assessment by educating assessment developers about aspects like assessment tasks, assessor training and obtaining satisfactory assessment results. The findings are organized under the familiar headings of validity, reliability, generalizability and educational consequences of assessments (Van der Vleuten 1996).

45.3.2.1 Findings with Regard to Validity

Validity refers to the extent to which a measurement (test, exam) measures what it is designed to measure. Three types of validity, i.e. content, criterion and construct validity, have long been distinguished, and in the classical view are conceptualized as intrinsic properties of a test. Construct validity refers to the test score as a measure of the assessed characteristic, which should be defined in a conceptual framework. Content validity focuses on the degree to which test content and response properties are representative of the domain in question. Criterion validity refers to the degree to which test scores predict future performance and correlate with results on other tests measuring the same construct.

Research findings on criterion validity in particular have yielded interesting results. For the *'knows how'* level, instruments were developed to measure candidates' reasoning and understanding, but studies in medical education revealed a strong correlation between complex paper-based patient scenarios and simple multiple-choice questions (Ward 1982; Swanson et al. 1987). This finding was contrary to the assumption that essays measured understanding and multiple-choice questions factual knowledge. Research on construct validity showed that information on content-specific knowledge and reasoning skills was difficult to generalize to other contents. With regard to content validity, candidates' responses to one assessment sample (question, case, situation, etc.) turned out to be poor predictors of performance on other samples, even within the same domain. This phenomenon was termed 'content specificity' or 'task variability' (Shavelson et al. 1993). These studies (Ward 1982; Swanson et al. 1987) created awareness that context and tasks, i.e. stimulus formats, had farther reaching consequences than did response formats (Van der Vleuten 1996). This insight made assessment developers realize that assessment tasks should present a faithful representation of the real workplace. This is in line with arguments from the 'authenticity movement' (Wiggins 1989; Cumming and Maxwell 1999), which promoted assessment in simulated or real-life authentic contexts. For the sake of authenticity, assessment tasks had to be pitched at the appropriate level of complexity, taking account of levels of cognitive functioning commensurate with a specific level of professional expertise (Van der Vleuten et al. 2010).

For the *'shows how'* level, live simulations were developed to differentiate between groups of candidates of diverse levels of experience (Van der Vleuten and Swanson 1990). Unfortunately, research showed that live simulations did not always discriminate between levels of expertise. An explanation for this was provided by research on expertise development showing that novices and experts differed not only in amount of knowledge but also in how they stored, used and retrieved knowledge (Schmidt et al. 1990). Professional expertise appeared to develop as a transition from a conceptually rich and rational knowledge base (acquired through educational experiences) to a non-analytical ability to recognize and handle situations efficiently and effectively (acquired through professional experiences). Such abilities were found to be difficult to transfer to other contexts. As it became increasingly clear that assessment was context dependent, it was realized that professional competence should preferably be assessed in authentic professional practice settings. Authenticity was of the essence for generalization from the measurement setting to other settings (Kane 1992). Kane introduced so-called 'high-fidelity tasks' (Kane 2006) for direct measurement of certain characteristics, which seemed best suited to the '*does*' level, i.e. practice-based or workplace-based assessments. Research has demonstrated that validity of this assessment depends strongly on how the assessor and the learner deal with the information that emerges from assessment, even more strongly than on the instruments used (Van der Vleuten et al. 2010). Assessors may have difficulty using scoring procedures and assessment criteria to interpret information from different contexts (Moss 1994), and candidates may strategically select information for inclusion in portfolios (Wolf and Dietz 1998). High-fidelity tasks, which are typically complex and open-ended, are hard for assessors to score (Kane 2006). This points to a need for assessors to be knowledgeable about assessment and trained to judge different sources of assessment information systematically and consistently, while candidates need to be informed about the purposes of assessments. The need for assessor training relates to reliability and generalizability, which will be discussed in the next section, while informing candidates relates to educational consequences, to which we will return later.

45.3.2.2 Findings with Regard to Reliability and Generalizability

Reliability relates to the replication of assessment results, i.e. the chance of finding different results when an assessment is repeated under the same conditions. Inter-rater reliability is often used as an indication of reliability (Dunbar et al. 1991). Traditionally, it has been assumed that assessors' judgements are more reliable when assessors consistently use carefully defined assessment criteria, performance levels and scoring rules (Moss 1994), whereas selective observations, personal prejudices and biases were considered to be serious threats to reliability and validity of assessment (Gipps 1994; Moss 1994). Assessor training is known to have a potentially positive impact on consistent scoring (Day and Sulsky 1995; Stamoulis and Hauenstein 1993), and global ratings are associated with a slight decrease in inter-rater reliability, while more analytical checklist scores yield higher inter-rater

reliabilities (Van der Vleuten et al. 2010). Reliability can also be improved by standardizing assessment tasks, for example by selecting tasks that represent key situations for a particular competence area. Standardization of tasks is also used to achieve generalizability, i.e. whether the sample of assessment tasks is representative of the universe of assessment tasks (Kane 1992), the collection of assessment tasks out of all possible tasks that are appropriate to measure the construct at hand.

Research has produced intriguing findings concerning the reliability and generalizability of methods for assessing the '*shows how*' and '*does*' levels of competence, with relevance to measures for scoring and selecting assessment tasks. Research on live simulations, unexpectedly revealed that compared to analytic judgements, global holistic judgements yielded better reliabilities across different tasks in live simulations (Rothman et al. 1997; Regehr et al. 1998). Apparently, global holistic ratings made judges more sensitive to elements in candidates' performance that were more generalizable across assessment tasks. Global ratings also resulted in scores that discriminated better between levels of expertise (Hodges et al. 1999; Norman 2005). A newer insight to emerge in relation to reliability is that it depends less on objectivity and standardization of methods and scoring procedures than on appropriate sampling of tasks and assessors (Kane 2006). When multiple assessors judge performance, threats to reliability, such as selective observation, biases and personal prejudices diminish, resulting in more accurate scoring. This implies that sampling across performances with different raters in each sample can considerably increase inter-rater reliability (Swanson 1987).

At the '*does*' level reliable scoring is considered to be a serious problem (Moss 1994), due to the variability of assessments and respondent reactions. Studies in medical education, investigating how direct observation, peer evaluations and multisource feedback impact on reliability, have yielded indications for the number of observations needed for adequate reliability (Kogan et al. 2009; Lockyer 2003; Falchikov and Goldfinch 2000; Davies et al. 2008; Moonen-Van Loon et al. 2013). Usually, a sample of 8–10 direct observations is sufficient, irrespective of the type of instrument and what is being measured, except for patient ratings which need larger samples. The discovery that a feasible sample of direct observations can produce adequate results fuelled assessment developers' enthusiasm for direct observation as a method for workplace-based assessment.

As for more indirect measures for the '*does*' level, such as portfolios, it is known that assessors struggle to consistently interpret materials from a variety of sources (Moss 1994), although moderately good inter-rater reliability has been shown to be achievable (Driessen et al. 2007). Bakker et al. (2011) showed that a clear and simple scoring procedure, with global criteria and discussions among raters, was effective provided raters were well prepared. The competence to be judged was stimulating and supporting self-regulated learning of students working collaboratively on complex tasks. The scoring procedure was based on a conceptual framework and considerations of situational awareness were included in the assessment criteria, which defined teachers as competent when they provided *just enough* support to enable students to move to the next level of learning, a move students would not have made successfully without teacher support (cf. Vygotsky 1978).

The assessors were trained to interpret video fragments of teacher performance in an authentic environment and to provide evidence and arguments for their judgements in accordance with the conceptual framework. Acceptable to high levels of inter-rater agreement were found, and assessors were reasonably competent to use the assessment procedure in a reliable manner. It took a substantial amount of training time, however, before teachers could recognize evidence in the video fragments and got used to the steps of the scoring procedure. In modern conceptions of reliability, elements of qualitative research are used to bring rigour to portfolio assessment (Driessen et al. 2005). Procedural measures provide evidence of due process in performance decisions, such as a specific number of feedback cycles before a summative decision is made, involvement of an independent committee, the number of assessors judging a single portfolio, the amount of justification provided for decisions, etc. As more of these measures are implemented, decisions become more trustworthy (Van der Vleuten et al. 2010).

45.3.2.3 Findings with Regard to Educational Consequences

The currently widespread notion that it is important to consider educational consequences of assessments is probably attributable to Messick's (1989) extended notion of validity, where validity is not only a test property but where the meaning or interpretation of scores must be valid as well as should any implications of that meaning. Incorporating consequential considerations into his definition of validity Messick proposed an integrated validity framework, combining issues of content, criterion and construct validity with considerations of value implications and social consequences to determine the impact of assessment on teaching and learning. This impact is often referred to as 'consequential validity', which also covers adverse effects of assessment, such as the use of undesirable learning strategies by students. Negative effects have frequently been attributed to knowledge tests targeted at Miller's *'knows'* level, because they were assumed to encourage rote learning aimed at reproducing knowledge without understanding, favouring a surface approach instead of a deep approach to learning (Biggs 1970, 1976; Entwistle and Entwistle 1970) or a reproduction-oriented learning style (Vermunt 1996) instead of a meaning-oriented learning style. Negative effects are known to be reinforced when tests are not judiciously distributed over the curriculum or compete with each other. Students may study very hard for a short time immediately before a test, and successfully reproduce the required knowledge at the test only to forget the knowledge as quickly as it was learned. Live simulations can generate meaningful information as well as enhance candidates' learning processes (Van der Vleuten and Swanson 1990), but the attending response formats can have adverse effects, as was illustrated in a study (Van Luijk et al. 1990) where students memorized detailed checklists and in their eagerness to show their 'knowledge' showed all behaviours on the list even when this was not appropriate for the situation at hand. The assessment developers quickly caught on to this phenomenon and switched to global rating scales adding to the criteria considerations of situational awareness.

For the '*does*' level evidence for the impact on learning is limited. Workplace-based assessment is used for assessment but also for its formative potential (Van der Vleuten et al. 2010). The provision of feedback to learners, in particular, gives workplace-based assessment formative value as it helps learners to steer their learning towards desired outcomes. Miller and Archer (2010) concluded that multisource or 360-degree feedback could improve performance, although personal factors, the feedback context and facilitation of feedback had a profound effect on candidates' responses to feedback. Feedback is most likely to bring about a change of performance when it is credible and accurate or when coaching is provided to help candidates identify and come to terms with their strengths and weaknesses. With regard to direct observation of procedural skills and case-based discussions, however, Miller and Archer found no evidence of improved performance concluding that further research was needed into the effects of workplace-based assessment on professional training and performance.

Research has examined the impact of more indirect methods of assessment at the '*does*' level. Driessen et al. (2007) showed that, for portfolios to successfully support learning and assessment, it is important that the goals and procedures are communicated clearly and portfolios are firmly embedded within the curriculum. Support from coaching or mentoring is prerequisite for effective support of learning. MacColgan and Blackwood (2009), however, found huge variation in types of portfolios for continuing professional development of educators and tremendous variability of terminology in portfolio research. Apparently, opinions differ regarding teaching portfolios and their uses complicating comparison of different studies on teaching portfolios. Comparisons of portfolios are also difficult because many studies report participants' perceptions without reporting measured effects on professional learning.

In recent years, there has been much research on assessment methods and procedures to promote self-directed and meaning-oriented learning. Formative assessment in particular is assumed to have a positive impact on student learning. Black and Wiliam (1998) analyzed findings from over 250 studies on formative assessment and concluded that the use of assessment outcomes to adjust learners' learning goals was a core characteristic (Black and Wiliam 1998 p. 140), i.e. evidence from formative assessment is used to give learners insight into where they stand in their learning, where they need to move and what they should do to get there (Black and Wiliam 2009). Wiliam and Thompson (2007) proposed five key strategies for formative assessment in classroom settings: (a) clarifying learning intentions and sharing criteria for success, (b) engineering effective classroom discussions, questions and learning tasks that elicit evidence of learning, (c) providing feedback that moves learners forward, (d) activating students as the owners of their own learning, and (e) activating students as instructional resources for each other.

However, implementing new modes of assessment does not automatically bring about desired changes in student learning, and various factors mediate the effects of new learning environments on student learning (e.g. Struyven et al. 2006). It is also a formidable challenge to deflect students' approaches to learning into a more desirable direction in a sustained way (e.g., Gijbels et al. 2008). Broekkamp and Van Hout-Wolters (2007) developed a theoretical model of factors which seem

relevant to adjusting learning strategies used by students to prepare for tests. This model may enhance our understanding of the different conditions and their interactions in this process. Based on ideas about strategy adaptation (Schunn and Reder 1998; Luwel et al. 2005) the model draws on considerations of teachers' test demands, students' perceptions of these demands, students' personal learning goals and students' ability to adjust and implement strategies. The framework acknowledges that strategy adaptation involves adaptation of external task characteristics, such as the environment where a task is undertaken, and internal task processes, such as students' ability to adjust learning strategies, their perceptions of task demands and the task disposition. Different results were found in a review of studies on adaptation of strategies for test preparation which varied across disciplines and between experimental and authentic study designs (Broekkamp and Van Hout-Wolters 2007). Although this should not be surprising, it underlines the relevance of their model and the need to systematically investigate factors that influence strategy adaptation in test preparation, both in experimental and authentic settings.

45.3.3 Summary of Implications

As we have shown, research into the quality of methods to assess professional competence has provided researchers and assessment developers with new insights with regard to the nature of professional competence and its assessment. We summarize the implications along the lines of a number of dimensions proposed by Segers (2004) (Fig. 45.2).

The first dimension refers to the change from decontextualized, atomized tests to authentic, contextualized assessment modes underpinned by insights that competence does not rely on generic, stable, independent traits but is context specific,

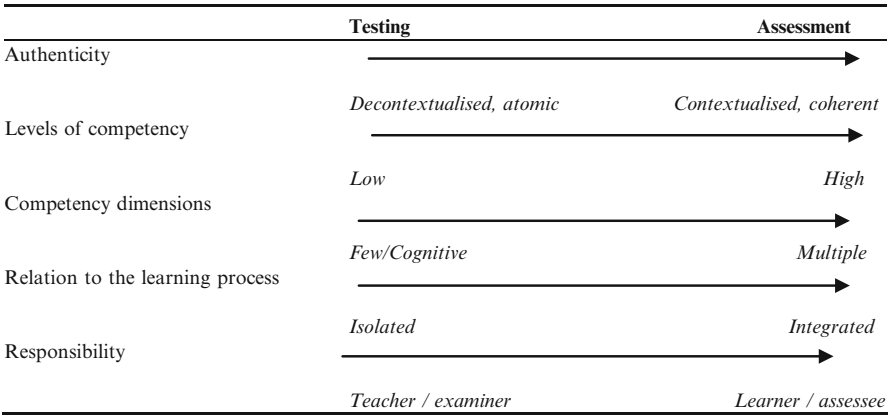


Fig. 45.2 Assessment characteristics and the differences between testing and assessment procedures as derived from Segers (2004, p. 9)

implying that account must be taken of the assessment context (Birenbaum 2003). Also large samples were needed for reliable and generalizable inferences about professional competence and information from various sources should be combined across content and time in order to gain a rich and multifaceted picture of a candidate's competence.

The second dimension is concerned with approaches to learning and professional development that emphasize lifelong learning (Atkins 1995; Eraut 1994). This involves the capturing of competency profiles that portray common patterns in the development of professional expertise using multiple methods and measurements over a prolonged period of time. This dimension is inspired by the insight that simulations may have shortcomings in differentiating between learners with different levels of expertise and can generate information that is difficult to generalize across contexts. To differentiate between expertise levels, assessment tasks should address levels of cognitive functioning that are representative of a certain level of professional expertise (Van der Vleuten et al. 2010), while scoring procedures should help assessors to distinguish levels of expertise, ranging from novice to expert. Assessment tasks and reference levels should support the development of professional expertise by capitalizing on self-directed and meaning-orientated learning.

The third dimension focuses on the multidimensionality of competencies including situational awareness as opposed to only knowledge and cognitive skills. Since assessment has evolved into a process where candidates show their competence in simulated or authentic contexts, assessment focuses on what candidates are able to and on actual performance in professional practice. This means that assessors should be able to determine to which extent candidates are responsive to what is appropriate in certain circumstances. This necessitates the inclusion of considerations of situational awareness in assessment criteria and preparation of assessors for working with these criteria, but on the other hand it leads to the acknowledgement that expert assessors are often already sensitive to these considerations.

The fourth dimension stresses the interconnectedness of assessment and learning, i.e. the notion that assessment drives learning (Longhurst and Norton 1997) implying that learning process and outcomes must be assessed in a relevant context (Dierick and Dochy 2001). It underscores the formative function of assessment, with feedback as a crucial factor (Sadler 1989), and requires monitoring of the impact of assessment on learning (Messick 1989). Recently, it has been proposed that assessment can and should be used strategically in assessment programmes to promote effective learning strategies and results (Van der Vleuten et al. 2010).

Finally, the fifth dimension relates to the responsibility for assessment, with a shift from control by examiners to assessee involvement in line with calls for learner participation in current views on learning (Birenbaum 2003). Assessee involvement is evident in the selection of items for inclusion in portfolios, but it can also be realized in 'tripartite meetings' where supervisor, assessor and student discuss the student's portfolio and the student submit additional information (e.g., Webb et al. 2003). Assessees can also be involved in the development and use of assessment criteria, such as in peer assessment (Sluijsmans et al. 2003). Negotiated assessment is an approach

that is particularly useful for promoting learning because of its participative and interactive elements (Gosling 2000; Boud 1992). It is characterized by extensive involvement of candidates in their own assessment and by an exchange of views between assessee and assessor, who are encouraged to negotiate and agree on the feedback provided and on the use of the assessment mechanism and criteria in light of learning objectives, activities and outcomes, based on the assumption that negotiations increase learner involvement and consequently enhance learning (Anderson et al. 1996). Assessors are expected to challenge learners who are reluctant to assume this active role (Anderson et al. 1996; De Eça 2005). Further research is needed, however, to determine its impact on learning (Verberg et al. 2013). In Table 45.1, along the lines of the dimensions we described and based on developments in assessment during recent decades, we summarize principles for the assessment of professional competence and its implications.

In the next section we will describe the move towards a programmatic approach to the assessment of professional competence.

Table 45.1 Research-based principles of assessment of professional competence

Dimension	Principle	(Practical) implications
(1) From decontextualized, atomized testing to authentic, contextualized assessments	Authenticity is key since competence is context specific (not generic)	Assessments should sample knowledge, skills and dispositions as used by professionals in professional practice Large samples to allow for reliable and generalizable inferences about candidates' competence Combinations of different assessment methods across content, time and assessment sources Assessment at 'does' level is reliable with 8–10 observations (by supervisor, peer, or multisource feedback); various stakeholders determine what candidates should know and be able to do Assessment tasks reflect the authentic context
(2) From low to high levels of competence	To distinguish different levels of development of professional competence, from novice to expert performance	Valid assessment requires tasks tailored to appropriate levels of complexity Reliable judgement requires appropriate scoring procedures and well-defined performance levels Valid results rely more on appropriate assessment tasks than on appropriate scoring procedures

(continued)

Table 45.1 (continued)

Dimension	Principle	(Practical) implications
(3) From cognitive dimensions of competency to multidimensional competency profiles	Professional competence requires situational awareness	Sensitivity of candidates and assessors to candidate's interactive cognitions and situational awareness relating to real situations in authentic contexts Aggregation of information from sources that are meaningfully similar (triangulation)
(4) From isolated tests to integration of assessment and learning	Assessment drives learning and can be used strategically as a learning tool	Combination of formative and summative functions on a continuum from low to high stakes assessment Evaluation for diagnosis and progress monitoring precedes final evaluations for certification or promotion Monitoring effects of assessments on learners Evaluations based on various sources of information Evaluations provide feedback for self-directed learning Preferably no combinations of multiple conflicting roles for assessors Combined design process for curriculum and assessment
(5) From control by assessors to shared control of assessors and assessees	Active involvement of candidates and/or sharing of control by assessors and candidates since self-directed learning supports the development of professional competence, active involvement and learner control	Extensive information for learners about purposes, requirements and procedures of assessment Learners responsible for providing information from themselves and others for the assessment Validity of assessment at 'does' level depends more on how assessors and learners use information from assessments than on particular assessment instruments Assessor training in judging, giving feedback and combining information from different sources – Learners can provide information (e.g., during intermediate assessment meetings) to supplement results of (practice-based) assessments that require interpretation of information from different sources

45.4 Programmatic Assessment of Professional Competence Fit for Purpose: A Model

Given the limitations of individual methods of assessment (Van der Vleuten 1996) a richer picture of a candidate's competence can be obtained by strategically combining multiple methods of assessment across content, time and assessment sources (Chester 2003; Van der Vleuten and Schuwirth 2005; Baartman et al. 2007; Van der Vleuten et al. 2012). Combining methods in an assessment programme means that modern approaches, such as live simulations, portfolios and practice-based assessment do not replace but rather supplement more traditional methods, such as knowledge tests. Careful selection of methods, formulation of rules and regulations and creating an organizational system can obtain a well-rounded picture of candidates' competence. According to Van der Vleuten and Schuwirth (2005) and Baartman et al. (2007) assessment is not a *psychometric problem* to be solved for one single method, but an *instructional design problem* encompassing the entire range of assessment methods used within the curriculum.

Fitness for purpose is the starting point for determining the quality of an assessment programme (Dijkstra et al. 2010, 2012). In a programmatic approach, all assessment purposes, including selection, monitoring and certification, are combined and optimized to maximize assessment for learning while at the same time arriving at sound decisions about learners' progress. Consequently, intermediate evaluations focused on diagnosis precede final evaluations, focused on high stakes decisions. All evaluations are based on various sources of information, and all information is aggregated to provide a sound basis for judgements, particularly for high-stakes decisions, which must be defensible (Van der Vleuten et al. 2010). Involvement of expert human judgement is considered to be imperative in assessment programmes because it is needed to judiciously combine assessment information to arrive at robust, defensible decisions at high stake moments and to tailor feedback to candidates' learning needs (Schuwirth and Van der Vleuten 2011). To counter threats to validity of scoring due to assessor-candidate relationships, assessors should be relieved from potentially compromising, multiple roles (Van der Vleuten et al. 2010). To enable high-stakes decisions based on aggregated information it is of the essence to prevent bias due to assessors having multiple roles.

In line with the notion of the integration of assessment and learning, assessment should have formative value to ensure its relevance to the learning process, and in a programmatic approach to assessment, formative and summative assessment functions are typically combined (Van der Vleuten et al. 2010). This means that assessment of competence in an authentic context should be given formative and summative weight (Van der Vleuten et al. 2010) to prevent that learners make strategic choices and do not take the assessment seriously thereby potentially trivializing the educational value of the assessment. Later in this chapter, we will elaborate on the combining of formative and summative assessment functions and some related dilemmas.

In programmatic approaches to assessment, qualitative, narrative information carries a lot of weight (Van der Vleuten et al. 2010), and assessment instruments have built-in facilities to elicit such information (e.g., space for narrative comments).

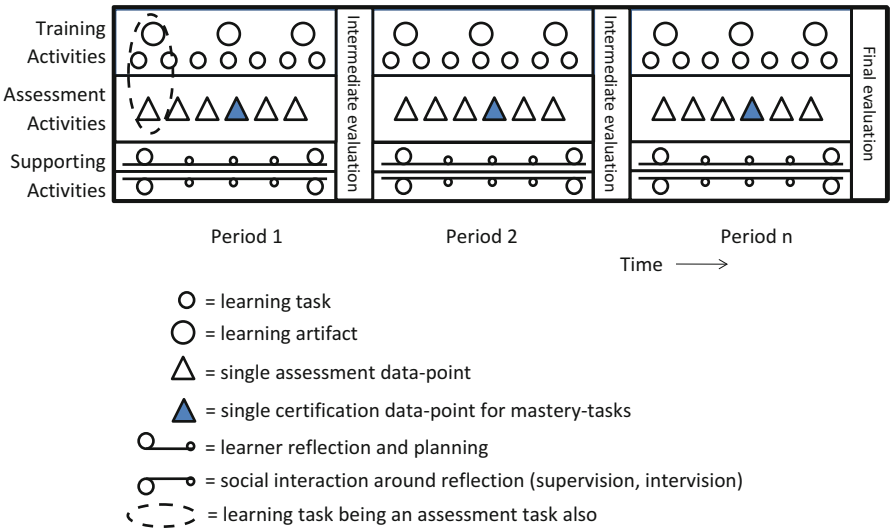
For reasons of transparency, this information has to be documented. All this has implications for assessor training: assessors should be trained to provide and document narrative feedback. Methodologies from qualitative research may support these processes (Tigelaar et al. 2005) and complement psychometric methods which are traditionally used to determine the quality of assessments. The same holds for monitoring on the programme level (Driessen et al. 2005). Later in this chapter, we will elaborate on assessment quality and related dilemmas.

Assessment programmes should be systematically evaluated for alignment with the curriculum and impact on learning, and evaluation results should be used to update the programme. Various stakeholders, including students, experts and practitioners should be involved in this process. This is important for meeting research-based and practice-based demands and expectations. However, as we will discuss later, integrating these demands in a productive way is often a struggle. In Box 45.1, we summarize the characteristics of assessment programmes.

Box 45.1 Characteristics of Assessment Programmes

An Assessment Programme…

- …. is grounded in a design that is based on an educational vision that supports both the curriculum and the assessment programme.
- … links competencies and assessment instruments in an overarching structure.
- … contains elements that produce information linked to specific courses or modules and elements that generate longitudinal information.
- … pays systematic attention to feedback, both quantitative and qualitative, to steer self-directed learning.
- … carefully balances formative and summative evaluations.
- … has panels for quality control implemented by (inter-collegial) test panels and committees.
- … is systematically evaluated with regard to alignment with the curriculum and impact on learning and uses the information from these evaluations to perform regular updates of the programme.
- … promotes active involvement of various stakeholders (students, teachers, and administrators) in the programme.
- …. contains intermediate evaluations for the purpose of diagnosis or progress monitoring, preceding final evaluations for certification or promotion.
- … draws on evaluations that are based on aggregated information from multiple sources, collected over time, and across time, content and different assessment sources.
- …has robust procedural arrangements to promote trustworthiness of the ultimate decisions about learners.



this figure will be printed in b/w

Fig. 45.3 A model for programmatic assessment. Reprinted from Van der Vleuten et al. 2012. A model for programmatic assessment fit for purpose. Medical Teacher, 34, 205-214, with permission from Informa Healthcare

Figure 45.3 visualizes a combination of different elements that may be part of an assessment programme. This model (Van der Vleuten et al. 2012) is based on the principles of assessment described in the previous section. Figure 45.3 shows that in the model we distinguish training activities, assessment activities and learner support activities as a function of the time in an ongoing curriculum. We illustrate the model with a case on teacher education.

45.4.1 Case: An Assessment Programme in Teacher Education

45.4.1.1 The Training Programme

The teacher education programme of the Leiden University Graduate School of Teaching, the Netherlands is a one year master programme aimed at training graduate students to become competent teachers. Students attend the programme full time or part time, and work as secondary school teachers, in an internship or part-time job, guided by a *school supervisor*, an experienced subject teacher specially trained for this task and a personal *university supervisor*, a lecturer who visits them at their school at least twice a year and observes them while teaching a class.

45.4.1.2 The Assessment Programme

The assessment programme addresses teaching competencies formulated in terms of teacher roles: a professional directing his/her own development, a subject teacher, a classroom manager, an adolescent psychologist, a member of the school organization and a researcher; rubrics are used to distinguish four levels of competence: poor, unsatisfactory, satisfactory and excellent. Tasks that teachers should be able to perform are described for each level.

Training/Learning Activities

Groups of student teachers and two supervisors are formed for the duration of the programme, and meet one day every week to discuss their experiences in teaching practice, (but student-supervisor conversation are also possible), attend focused courses on topics such as subject methodology, psychology of adolescents, classroom management and preparation of a research proposal. The remaining time is devoted to preparation for learning activities and assignments which may yield *artefacts of learning*, such as planning a lesson, evaluating the results, selected readings on relevant topics, constructing a sociogramme (a chart showing the natural friendship relations in a class) and administering and analyzing a questionnaire on teacher interaction (QTI, Wubbels and Levy 1991) in the students' own classrooms.

Assessment Activities

Single data points of assessment produce low-stakes information related to courses or modules, such as a video recording of a lesson followed by group discussion or artefacts, such as the analysis of a QTI. Feedback from supervisors and peers is provided in a global rating format, with comments and suggestions aimed at maximizing the impact on learning. Each data point is one element in a longitudinal array of data points.

Supporting Activities

The supporting activities focus on feedback to guide students' self-directed learning in terms of *learner reflection and planning*. Feedback on learning and assessment tasks, such as the analysis of the QTI or a video recording of a lesson, is interpreted and used to set new learning goals (Van Merriënboer and Sluijsmans 2009). To scaffold self-directed learning social interactions are arranged, such as the university supervisor asking questions to elicit reflective activities, such as describing, analyzing and planning. The student-teachers receive training for and engage in collaborative reflections with peer students.

Intermediate Evaluations

Intermediate evaluations are scheduled at the end of each module, on completion of the practical experience and in two formal assessment meetings: the go/no go assessment and the midterm evaluation. The former evaluates performance in the roles of subject specialist, classroom manager and professional, based on the student's self-evaluation and a global judgement of the university and school supervisor. The aim is s*election*, i.e. students who are considered to be unlikely to become successful teachers are told this early in the course (i.e. after 2 months) to prevent disappointment later on. Students can disregard this advice, but they are not entitled to extra support if they run into difficulties. If there is doubt about a student's prospects, a tripartite meeting is organized in which the supervisors and the student discuss the student's prospects.

The diagnostic intermediate evaluation assesses a student's progress in all roles based on: a video, which the student analyzes for classroom management, a case analysis on adolescent psychology based on the student's teaching practice and theoretical notions, a series of four lessons, analyzed for subject methodology, a cased-based examination of theories of learning and instruction and their applications (written test), a self-analysis of the student's functioning in the school context, with multisource feedback (from pupils, the school supervisor, a fellow teacher of the same subject and a school leader).

Final Evaluations

The prime aim of the final assessment of student performance in the six roles is certification: has the student successfully completed the required modules and presented a complete portfolio to the assessors. The assessment can be diagnostic too, as it may involve a discussion with the student about their plans after the course. The complete portfolio contains: four lessons to be judged on subject methodology, consisting of lesson plans and assessment instruments, including a theoretical underpinning of individual lesson plans and how the lessons are connected, evaluation results of pupils, a video of a lesson, analyzed by the student with suggestions for improvement, a 5,000 word paper reporting on the student's research project, and an analysis of the student's performance in all roles, based on theoretical notions and a self-analysis of the student with multisource feedback (from pupils, the supervisor in the school, a colleague teacher of the same subject and a school leader) on the different roles. Judgements are made using (global) rating scales.

People involved in the assessment programme.

- The *supervisor is responsible for the different modules* and judges whether the student has met the demands of a specific module. The supervisors are the subject specialist and the supervisor on adolescent psychology.

- The *school supervisor and/or job coach* is co-assessor of the final assessment, and advises on intermediate and final assessments in relation to the practical assessment. The supervisor of the school where the student teachers is involved in the final assessment.
- The university *supervisor* is responsible for the go/no go-assessment, the midcourse evaluation and, together with the school supervisor, the final assessment. The university supervisor seeks further information and advice from others in making these assessments.
- The *examination committee:* in cases of disagreement about assessment between student and supervisors or between supervisors, the examination committee has the final word.

45.4.1.3 Some Evaluative Remarks on the Assessment Programme for Student Teachers

The model of programmatic assessment we have just described meets the characteristics of assessment programmes summarized in Box 45.1, but evaluations have indicated that improvements are possible. The single data points are not always clearly linked to the overarching aims of the programme while the purpose of the assignments is not always clear to students. Some students feel that self-analysis and the research assignment have no value for their development as a teacher. To maximize learning and support sound judgement processes, the assessment instruments and the way they are used for formative and summative purposes needs careful re-evaluation in light of the educational vision underlying the programme. Another area for improvement is the quality of judgements: school and university supervisors sometimes differ in their judgements of students' teaching competence. This may be due to difficulties in assessing different types of information or university and school supervisors, despite having participated in the development of the programme, disagreeing about the requirements for adequate functioning as a professional in teaching practice. Discussions with various stakeholders in the assessment programme may be necessary to determine which types of information are needed for judgements and to advance the meaningfulness of the rubrics and scoring procedures. Assessor training deserves more attention and may be improved by using concrete exemplars from professional practice.

The teacher education programme we have described is one illustration of an assessment programme. Descriptions of other assessment programmes designed according to the same model (Dannefer and Henson 2007; Driessen et al. 2012; Altahawi et al. 2012) show that designers of assessment programmes should always keep in mind that passing examinations and university tests may meet the needs of institutions but that ultimately the worth and competence of graduates is appraised in other ways and that students will use different standards to evaluate their preparation for teaching practice. This means that the link between assessment for educational and practical purposes is a crucial one. In the next section, we discuss some issues

related to that link. In the final section of this chapter, we describe additional research findings and discuss future prospects with regard to guidelines for designing assessment programmes.

45.5 Issues in the Assessment of Professional Competence

This section outlines issues with regard to linkage of assessment for educational and for practice purposes, dilemmas in combining formative and summative assessment and monitoring quality in the assessment of professional competence.

45.5.1 Linking Assessment for Educational and Practical Purposes

Issues that arise in relation to linking educational and practical purposes of assessment often revolve around incompatibility of different stakeholders' ideas about the definition of professional competence in an authentic context. Although the literature indicates that criteria and standards for professional competence should be based on empirical evidence, cooperation with practitioners on criteria and standards is also important, particularly for areas of professional competence with a relatively recent knowledge base, such as teaching (Uhlenbeck et al. 2002). Theoretical notions with regard to the acquisition of competence within a particular area need to be translated into a language that reflects the experiences and problems which professionals face in daily practice. This is important not only for fostering a sense of ownership but also, and perhaps even more so, to ensure that assessments are meaningful to daily practice and acceptable. Although professional may regard evidence-based theoretical notions as supportive to expertise development (Van Driel and Berry 2010), research also describes instances where professionals make hardly any use of theoretical notions on their own field of practice. In assessment it seems crucial that in designing assessment programmes learners' experiences and practical concerns are taken as the point of departure. An example of how this can be done, while also aiming to acquaint students with theoretical notions and evidence-based findings, can be found in a paper by Nilsson (2013) showing how research techniques can be used as formative assessment tools for developing primary science student teachers' pedagogical content knowledge (PCK) and for assisting them in becoming aware of their own PCK in relation to their own teaching. Formative assessment consisted of activities by teacher educators to stimulate interactions and self- and peer-assessment in order to provide insights into how student teachers develop their PCK during a semester. A research tool, Content Representations (CoRes), was used to unpack student teachers' approach to teaching a science topic and stimulated recall seminars using video recordings of lessons given by students were used to encourage formative

interaction between the student teachers and the teacher educator. The CoRes were used to measure student-teachers' PCK and have them reflect on their PCK, and may be seen as implicit reference levels, describing aspects of performance that need to be assessed (i.e., criteria) and portraying the student teachers' *own* development. The results of Nilsson's exploration indicate that the use of CoRes, together with subsequent self-assessment and formative interactions with teacher educators and peers were considered as relevant by the student teachers and at the same time have potential for PCK development for student teachers.

This example illustrates how theoretical notions and concepts, such as PCK development of science teachers, can be used productively to improve coherence between notions used in education and the demands of practice when it comes to adequate professional competence. Discussing findings from research with practitioners in the field may also be helpful to improve integration in this respect. Research findings can provide useful input for constructing authentic cases that provide exemplars of different levels of professional performance in relation to evidence-based standards. Such exemplars can be helpful for reaching shared understanding between assessors from educational institutes and schools who have to judge student performances, but also for making candidates sensitive to what is expected of them as professionals.

45.5.2 Combining Formative and Summative Assessment

Since it is known that assessment drives learning (Longhurst and Norton 1997), assessment processes should be designed to provide meaningful learning experiences and give candidates a fair chance of displaying their competence. However, there is the threat of undesirable interference of these two goals which may not be compatible. There may be negative backwash effects on candidates' learning processes when candidates influenced by considerations of summative assessment exclusively present their strong points (Biggs 1996, 1999). Similarly, in the context of portfolio assessment, candidates may become very selective in including items in their portfolios and in writing comments on their teaching performance. As a result the formative function of assessment is reduced since areas where improvement is needed remain underexposed to feedback and critical reflection. Nevertheless, we think that formative and summative assessments should be integrated wherever possible, and this is supported by a study showing that the ability to work towards a summative decision from the start motivated teachers to work on their portfolio (Tigelaar et al. 2006). The key issue here is that Apparently, it is not realistic to expect intrinsic motivation to offer a sufficient incentive for candidates to spend time on their portfolios and on reflecting on their teaching performance. Candidates also need to be convinced that the ultimate goals and profits make their efforts worthwhile.

The question of who is responsible for assessing a candidate's progress and providing support and feedback is another dilemma. There is a potential conflict of

interest when supervisors and mentors are required to combine the roles of guide of learning and judge of competence achievement (Tigelaar and Van Tartwijk 2010). This dilemma may be resolved by considering different assessment scenarios (Van Tartwijk et al. 2003), such as the "job-application scenario", where the (committee) of assessors is independent and the candidates are responsible for preparing the assessment information, without any guidance or supervision. Another scenario is the "driving exam" scenario, where a supervisor helps candidates to prepare for the assessment, and an independent assessor assesses the candidate's competence, without consulting the supervisor. Finally, in the "PhD supervisor" scenario, the supervisor helps the candidate attain the required level of competence, and decides that the collected evidence can be submitted to an assessment committee (like a professor supervising a PhD thesis). Most of the time the committee will confirm the supervisor's decision. But if the supervisor does not do a proper job and is too lenient, the assessment committee may reach a negative conclusion.

The dilemma can also be resolved by limited involvement of coaches in the assessment of their own 'pupils'. In order to overcome threats to the validity of scoring due to a relationship between assessor and candidate, it is argued that assessors need to be relieved from potentially compromising, multiple roles (Van der Vleuten et al. 2010). In high-stakes decisions based on aggregated information, procedures are needed to prevent bias in assessment processes caused by assessors having multiple roles. An example of such a procedure is described by Driessen et al. (2005), who argue that supervisors should not be the formal assessors of the candidates they support and guide because they are too closely involved with them. In this approach supervisors give feedback to the candidates they have guided before candidates submit their portfolio to the assessment committee, but the candidates are responsible for presenting their portfolio to the assessment committee. This is in line with constructivist views on learning and assessment, which stress learner participation and control (Birenbaum 2003; Segers 2004). The coach of the teacher who is being assessed may provide additional context information as a member of the team of assessors, which otherwise consists of other members, who are knowledgeable on teaching from different perspectives.

45.5.3 Quality in the Assessment of Professional Competence

In this section, we elaborate on methods to monitor and support assessment quality, and dilemmas inherent in this enterprise.

Earlier we discussed various psychometric criteria and we stated that assessors' judgements are more reliable when assessors use carefully defined assessment criteria, performance levels and scoring rules in a consistent manner (Moss 1994). This may explain why assessor training can have a positive influence on the consistent application of criteria, standards and scoring rules (Day and Sulsky 1995; Stamoulis and Hauenstein 1993). A well-defined assessment framework is required from which criteria, standards and scoring roles can be derived. The framework contains

description of the constructs of professional competence to be assessed, such as surgical/procedural skills and respect for patients and other aspects of the professional competence of medical doctors (Messick 1989). Asking assessors to give rationales for their judgements (Baume et al. 2004) may help to refine the description of constructs and the criteria, standards and rules for scoring. This is consistent with our earlier argument that expert human assessors may bring additional insights to assessments which move beyond merely adhering to scoring rules and may add to the validity of the assessment. As a consequence, a more instrumental approach to defining and using assessment frameworks may be adopted, providing room for multiple perspectives on what is important in a certain area of professional competence, and putting more weight on expert human judgement (Moss 1994). In such an approach, assessment results can still be used to inform conceptualizations of professional competence and vice versa. Apart from various aspects of construct validity, other aspects of assessment quality can be considered to complement psychometric approaches to determining assessment quality, such as approaches inspired by methodologies for establishing credibility in qualitative research (Tigelaar et al. 2005; Driessen et al. 2005). Instead of focusing on standardization of scoring procedures and banning influences from assessors' idiosyncratic frames of reference, such methodologies can support the quality of scoring by stimulating assessors to give rationales for their judgements, by documenting assessment processes so that they can be made available to candidates thereby making the assessment process more meaningful for supporting learning and to enable others to check the conclusions of the assessment.

45.6 Future Prospects

This chapter ends with a number of future prospects, summarized under the headings: supporting expert judgement, developing guidelines for assessment programmes and gaining insight into underlying mechanisms that may explain the impact of assessment on learning.

45.6.1 Supporting Expert Judgement

As said earlier, expert professional judgement is imperative, especially in authentic assessment in an assessment programme. We have also argued that psychometric approaches to assessment quality should be complemented with other approaches, such as integration of assessment in instructional design and assessment methodologies based on notions from qualitative research. We need to know more about assessors' reasoning processes to prevent these methods from becoming trivialized as well. Although recent research into assessors' reasoning processes has provided valuable insights into characteristics of these processes and practical implications for supporting assessors' scoring processes (Schutz and Moss 2004; Govaerts et al. 2007, 2011;

Bakker et al. 2011), more research is needed on how expert judgement can be supported in a productive way. A balance may have to be struck between approaches that focus on standardization of assessment procedures and approaches following a more open procedure with extensive documentation of the judgement process. This holds not only for decisions on single point assessments, but even more for high-stakes decisions about selection and certification of candidates in assessment programmes. Since it is crucial for high-stakes decisions to be credible and defensible, more research is needed on appraisal of all the relevant evidence collected in an assessment programme in a sound, transparent and meaningful way. Since psychometric objectification and standardization tend to trivialize the assessment process (Van der Vleuten et al. 1991), we might look to qualitative research methodologies but also to research on naturalistic decision making (Klein 2008) and law (Simon 2004). Research on naturalistic decision making has shown that people rarely employ systematic or algorithmic strategies (Kahneman et al. 1982), but use prior experience and intuition (Dijksterhuis et al. 2006). Research in law has addressed decision-making 'on the balance of probabilities' or 'beyond reasonable doubt' (Simon 2004), showing that assessors' mental representations shift while weighing evidence until their reasoning very strongly and coherently points in a certain direction. These theories may prove to be worthwhile for assessment programmes without placing a heavy burden on assessors.

45.6.2 Developing Guidelines for Assessment Programmes

By way of illustration, we presented an assessment programme in a graduate teacher training course. Other assessment programmes have been described in the literature (e.g., Dannefer et al. 2007; Driessen et al. 2012). The differences between programmes are partly related to what is known about professional competence in a particular field in terms of construct definitions and methods which, in a deliberate arrangement of activities, may be useful to include in an assessment programme to obtain a well-rounded picture of a candidate's competence and to steer candidates' learning processes. We know much more about professional competence in the context of medical education and about the development of expertise in particular domains of medical expertise than in many other domains. Another reason why assessment programmes can take different forms is determined by the context. Since the actual requirements for professional practice differ widely across circumstances, we agree with Kane (1992) that the requirements of real practice which future professionals have to meet should always be taken into account when designing an assessment programme. An assessment programme that lasts only one year may include fewer high-stakes decisions. As a consequence, measures to account for the quality of assessment decisions may be less complex. There are other aspects that may explain differences between assessment programmes and how they are constructed, but the key starting point should be the purpose of the programme. We already indicated that, although

often multiple purposes relating to selection, diagnosis and certification may be combined, it is usually one purpose that receives special emphasis (Hickey et al. 2006). We also contended that *fitness for purpose* of an assessment programme should be the starting point when determining its quality (Dijkstra et al. 2010). This implies that guidelines for building assessment programmes are best formulated generically in order to ensure applicability to various contexts. Recently, Dijkstra et al. (2012) conducted a study with a number of experts in assessment to validate fit-for-purpose guidelines for designing programmes of assessment. Their study resulted in a set of guidelines that is comprehensive and not bound to specific contexts or educational approaches. Among these guidelines are generic guiding principles focused on the importance of underpinning decisions in assessment programmes by collecting, combining and valuing information and taking concrete actions. Other guidelines point to ways to support the assessment programme and to use documentation in a productive way for improving the programme. The guidelines are formulated eclectically, which means that they rely on professional judgement for appropriate use in a particular context. Although further analysis of assessment programmes in various contexts is necessary to validate these guidelines, the available guidelines, combined with what is already known about programmes of assessment may enable assessment designers in various areas of professional education and training to monitor the complex dynamics of programmatic assessment in their own context.

45.6.3 Mechanisms Underlying the Impact of Assessment on Learning

Inspired by the insight that assessment drives learning, we have argued that assessment should be used strategically to monitor and support learning. However, we also mentioned the need for more knowledge about mechanisms underlying the impact of assessment on learning. Earlier in this chapter, we mentioned some studies that provide valuable anchor points for further research with regard to this impact. Cilliers et al. (2010, 2012) recently added to this field of research by studying such mechanisms in the context of summative assessment. They explored mechanisms underlying the impact of assessment on learning in an in-depth interview study. Mechanisms that emerged from the analysis of the results were the ways students appraise the impact of various assessment methods in a curriculum, their own learning response, their own perceptions of agency and contextual factors. Task demands, imminence of the assessment, the design of the assessment system and the cues, inferred from the assessors or the assessment tasks, that informed students on what content to learn, emerged as factors that determined the impact. Cognitive and meta-cognitive regulation activities emerged as consequences of assessment for learning (Cilliers et al. 2012). These findings are helpful for improving our understanding of the mechanisms in

assessment that drive learning. Studies, such as those by Cilliers et al. (2010, 2012) and Bennett (2011), have given an impetus to further develop theoretical models that explain this impact. This type of research should be continued to understand mechanisms in the assessment of professional competence and their impact on how future professionals learn and develop their professional competence.

References

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.

Anderson, P., & Lawton, L. (1988). Assessing student performance in a business simulation exercise. *Developments in Business Simulation & Experiential Exercises, 15*, 241–245.

Anderson, G., Boud, D., & Sampson, J. (1996). *Learning contracts*. London/New York: Routledge Farmer.

Andrews, T. E., & Barnes, S. (1990). Assessment of teaching. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 569–598). New York: Macmillan.

Atkins, M. (1995). What should we be assessing. In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 25–33). London: Kogan Page.

Baars, B. J. (1986). *The cognitive revolution in psychology*. New York: The Guilford Press.

Baartman, L. K. J., Bastiaens, T., Kirschner, P. A., & Van Vleuten, C. P. M. (2007). Teachers' Opinions on quality criteria for competency assessment programs. *Teaching and Teacher Education, 23*(6), 857–867.

Bakker, M., Roelofs, E., Beijaard, D., Sanders, P., Tigelaar, D., & Verloop, N. (2011). Video portfolios: The development and usefulness of a teacher assessment procedure. *Studies in Educational Evaluation, 37*, 123–133.

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*(4), 451–477.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Biggs, J. B. (1970). Faculty patterns in study behaviour. *Australian Journal of Psychology, 22*, 161–174.

Biggs, J. B. (1976). Dimensions of study behaviour: Another look at ATI. *British Journal of Educational Psychology, 46*, 68–80.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364.

Biggs, J. B. (1999). *What the student does: Teaching for quality learning at university*. Buckingham: Open University Press.

Bird, T. (1990). The schoolteacher's portfolio: An essay on possibilities. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 241–256). Newbury Park: Corwin Press.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 13–36). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

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

Boud, D. (1992). The use of self-assessment schedules in negotiated learning. *Studies in Higher Education, 17*(2), 185–200.

Broekkamp, H., & Van Hout-Wolters, B. H. A. M. (2007). Students' adaptation of study strategies when preparing for classroom tests. *Educational Psychology Review, 19*(4), 401–428.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*, 32–41.

Cilliers, F. J., Schuwirth, L. W. T., Adendorff, H. J., Herman, N., & Van der Vleuten, C. P. M. (2010). The mechanisms of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education: Theory and Practice, 15*(5), 695–715.

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. N., & Van der Vleuten, C. P. M. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education: Theory and Practice, 17*(1), 39–53.

Cumming, J., & Maxwell, G. (1999). Contextualising authentic assessment. *Assessment in Higher Education, 6*(2), 177–194.

Dannefer, E. F., & Henson, L. C. (2007). The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine: Journal of the Association of American Medical Colleges, 82*(5), 493–502.

Dannefer, E. F., Henson, L. C., & Lindsey, C. (2007). The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine, 82*(5), 493–502.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*, 285–328.

Davies, H., Archer, J., Bateman, A., Dewar, S., Crossley, J., Grant, J., & Southgate, L. (2008). Specialty-specific multi-source feedback: Assuring validity, informing training. *Medical Education, 42*(10), 1014–1020.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158–167.

De Eça, M. T. T. P. (2005). Using portfolios for external assessment: An experiment in Portugal. *International Journal of Art & Design Education, 24*(2), 209–218.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307–329.

Dijksterhuis, A., Bos, M. W., Nordgren, L. T., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*, 1005–1007.

Dijkstra, J., Van der Vleuten, C. P. M., & Schuwirth, L. W. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education: Theory and Practice, 15*, 379–393.

Dijkstra, J., Galbraith, R., et al. (2012). Expert validation of fit-for-purpose guidelines for designing programs of assessment. *British Medical Education, 12*(1), 20.

Driessen, E. W., Van der Vleuten, C. P. M., Schuwirth, L. W., 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*(2), 214–220.

Driessen, E. W., Van Tarwijk, J., Van der Vleuten, C. P. M., & Wass, V. (2007). Portfolios in medical education: Why do they meet with mixed success? A systematic review. *Medical Education, 41*(12), 1224–1233.

Driessen, E. W., Van Tarwijk, J., Govaerts, M., Teunissen, P., & Van der Vleuten, C. P. M. (2012). The use of programmatic assessment in the clinical workplace: A Maastricht case report. *Medical Teacher, 34*(3), 226–231.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289–303.

Entwistle, N. J., & Entwistle, D. (1970). The relationships between personality, study methods and academic performance. *British Journal of Educational Psychology, 40*(2), 132–143.

Eraut, M. (1994). *Developing professional knowledge and competence*. London: The Falmer Press.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

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

Gijbels, D., Segers, M., & Struyf, E. (2008). Constructivist learning environments and the (im) possibility to change students' perceptions of assessment demands and approaches to learning. *Instructional Science, 36*(5–6), 431–443.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London/ Washington, DC: The Falmer Press.

Gosling, D. (2000). Using Habermas to evaluate two approaches to negotiated assessment. *Assessment & Evaluation in Higher Education, 25*(3), 293–304.

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

Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education: Theory and Practice, 16*, 151–165.

Haertel, E. (1991). New forms of teacher assessment. In G. Grant (Ed.), *Review of research in education* (Vol. 17, pp. 3–29). Washington, DC: American Educational Research Association.

Harden, R., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*, 41–54.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation, 32*, 180–201.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74*(10), 1129–1134.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation & the Health Professions, 15*(2), 163–182.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Education measurement* (4th ed.). Westport: Praeger Publishers.

Klein, G. (2008). Naturalistic decision-making. *Human Factors, 50*, 456–460.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*(12), 1316–1326.

Lockyer, J. (2003). Multisource feedback in the assessment of physician competencies. *Journal of Continuing Education in the Health Professions, 23*(1), 4–12.

Longhurst, N., & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*, 319–330.

Luwel, K., Lemaire, P., & Verschaffel, L. (2005). Children's strategies in numerosity judgment. *Cognitive Development, 20*(3), 448–471.

MacColgan, K., & Blackwood, B. (2009). A systematic review protocol on the use of teaching portfolios for educators in further and higher education. *Journal of Advanced Nursing, 65*(12), 2500–2507.

Mansvelder-Longayroux, D. D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching & Teacher Education, 23*(1), 47–62.

McGuire, C. H., & Babnott, D. (1967). Simulation technique in the measurement of problem solving skills. *Journal of Educational Measurement, 4*, 1–10.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: MacMillan.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), 63–67.

Miller, A., & Archer, J. (2010). Impact of workplace based assessment on doctors' education and performance: A systematic review. *British Medical Journal, 341*, c5064. doi:10.1136/bmj.c5064.

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., Van der Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087–1102. doi:10.1007/s10459-013-9450-z.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5–12.
Nilsson, P. (2013). What do we know and where do we go? Formative assessment in developing student teachers' professional learning of teaching science. *Teachers and Teaching: Theory and Practice, 19*(2), 188–201. doi:10.1080/13540602.2013.741838.
Norman, G. (2005). Editorial-Checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Advances in Health Sciences Education: Theory and Practice, 10*(1), 1–3.
Regehr, G., MacRae, H., Reznick, R., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*(9), 993–997.
Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal for Vocational Training, 40*(1), 123–139.
Rothman, A. I., Blackmore, D., Dauphinee, W. D., & Reznick, R. (1997). The use of global ratings in OSCE station scores. *Advances in Health Sciences Education: Theory and Practice, 1*, 215–219.
Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.
Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: Theory and implication [published erratum appears in Academic Medicine, 1992, 67(4): 287]. *Academic Medicine*, *65*(10), 611–621.
Schunn, C. D., & Reder, L. M. (1998). Strategy adaptivity and individual differences. *Psychology of Learning and Motivation, 38*, 115–154.
Schutz, A. M., & Moss, P. A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives, 12*(33). Retrieved July 19, 2004, from http://epaa.asu.edu/v12n33/
Schuwirth, L. W., & Van der Vleuten, C. P. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33*(6), 478–485.
Segers, M. (2004). *Assessment en leren als een twee-eenheid: Onderzoek naar de impact van assessment op leren.* [The dyad of assessment and learning: A study of the impact of assessment on learning]. Inaugural address given at the acceptance of professorship in Pedagogics, Educational Sciences in particular, at Leiden University. Leiden: Leiden University.
Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.
Shulman, L. S., Haertel, E. H., & Bird, T. (1988). *Toward alternative assessments of teaching. A report of work in progress*. Palo Alto: Stanford University.
Simon, D. (2004). A third view of the black box. Cognitive coherence in legal decision-making. *The University of Chicago Law Review, 511*, 512–513.
Sluijsmans, D. M. A., Brand-Gruwel, S., Van Merriënboer, J., & Bastiaens, T. R. (2003). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29*, 23–42.
Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rater accuracy: Training for dimensional accuracy versus training for rater differentiation. *Journal of Applied Psychology, 78*(6), 994–1003.
Stoof, A., Martens, R., Van Merriënboer, J., & Bastiaens, T. (2002). The boundary approach of competence: A constructivist aid for understanding and using the concept of competence. *Human Resource Development Review, 1*, 345–365.
Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2006). On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and Instruction, 16*(4), 279–294.
Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 13–45). Montreal: Can-Heal publications.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*(3), 220–246.

Tigelaar, D. E. H., & van Tartwijk, J. (2010). The evaluation of prospective teachers in teacher education. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 7, pp. 511–517). Oxford: Elsevier.

Tigelaar, D. E. H., Dolmans, D. H. J. M., Wolfhagen, H. A. P., & Van der Vleuten, C. P. M. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*(5), 595–610.

Tigelaar, D. E. H., Dolmans, D. H. J. M., De Grave, W. S., Wolfhagen, H. A. P., & Van der Vleuten, C. P. M. (2006). Participants' opinions on the usefulness of a teaching portfolio. *Medical Education, 40*(4), 371–378.

Uhlenbeck, A., Verloop, N., & Beijaard, D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *The Teachers College Record, 104*(2), 242–272.

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

Van der Vleuten, C. P. M., & Schuwirth, L. T. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*, 309–317.

Van der Vleuten, C. P. M., & Swanson, D. (1990). Assessment of clinical skills with standardised patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58–76.

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

Van der Vleuten, C. P. M., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics and Gynaecology, 24*, 703–719.

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

Van Driel, J. H., & Berry, A. (2010). The teacher education knowledge base: Pedagogical content knowledge. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 7, pp. 656–661). Oxford: Elsevier.

Van Luijk, S. J., Van der Vleuten, C. P. M., & Van Schelven, R. M. (1990). The relation between content and psychometric characteristics in performance-based testing. In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, et al. (Eds.), *Teaching and assessing clinical competence* (pp. 202–207). Groningen: Boekwerk Publications.

Van Merriënboer, J. G., & Sluijsmans, M. A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed Learning. *Educational Psychology Review, 21*, 55–66.

Van Tartwijk, J., Driessen, E. W., Hoeberigs, B., Kösters, J., Ritzen, M., Stokking, K., & Van der Vleuten, C. P. M. (2003). *Werken met een electronisch portfolio* [Working with an electronic portfolio]. Groningen: Wolters Noordhoff.

Van Tartwijk, J., Driessen, E., van der Vleuten, C., & Stokking, K. M. (2007). Factors influencing the successful introduction of portfolios. *Quality in Higher Education, 13*(1), 69–79.

Verberg, C. P. M., Tigelaar, D. E. H., & Verloop, N. (2013). Teacher learning through participation in a negotiated assessment procedure. *Teachers and Teaching, 19*(2), 172–187. doi:10.1080/13540602.2013.741842.

Vermunt, J. D. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education, 31*(1), 25–50.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University press.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6*(1), 1–11.

Webb, C., Endacott, R., Gray, M. A., Jasper, M. A., McMullan, M., & Scholes, J. (2003). Evaluating portfolio assessment systems: What are the appropriate criteria? *Nurse Education Today, 23*(8), 600.
Wiggins, G. (1989). A true test. *Phi Delta Kappan, 70*(9), 703–713.
Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah: Lawrence Erlbaum Associates.
Wolf, K., & Dietz, M. (1998). Teaching portfolios; Purposes and possibilities. *Teacher Education Quarterly, 25*(1), 9–22.
Wubbels, T., & Levy, J. (1991). A comparison of interpersonal behavior of Dutch and American teachers. *International Journal of Intercultural Relations, 15*, 1–18.