Adv in Health Sci Educ (2013) 18:1009–1027
DOI 10.1007/s10459-013-9444-x

An argument-based approach to the validation of UHTRUST: can we measure how recent graduates can be trusted with unfamiliar tasks?

M. Wijnen-Meijer · M. Van der Schaaf · E. Booij · S. Harendza · C. Boscardin · J. Van Wijngaarden · Th. J. Ten Cate

Received: 10 August 2012 / Accepted: 16 January 2013 / Published online: 12 February 2013
© Springer Science+Business Media Dordrecht 2013

Abstract There is a need for valid methods to assess the readiness for clinical practice of medical graduates. This study evaluates the validity of Utrecht Hamburg Trainee Responsibility for Unfamiliar Situations Test (UHTRUST), an authentic simulation procedure to assess whether medical trainees are ready to be entrusted with unfamiliar clinical tasks near the highest level of Miller's pyramid. This assessment, in which candidates were judged by clinicians, nurses and standardized patients, addresses the question: can this trainee be trusted with unfamiliar clinical tasks? The aim of this paper is to provide a validity argument for this assessment procedure. We collected data from various sources during preparation and administration of a UHTRUST-assessment. In total, 60 candidates (30 from the Netherlands and 30 from Germany) participated. To provide a validity argument for the UHTRUST-assessment, we followed Kane's argument-based approach for validation. All available data were used to design a coherent and plausible argument. Considerable data was collected during the development of the assessment procedure. In addition, a generalizability study was conducted to evaluate the reliability of the scores given by assessors and to determine the proportion of variance accounted by candidates and assessors. It was found that most of Kane's validity assumptions were defendable with accurate and often parallel lines of backing. UHTRUST can be used to compare the

M. Wijnen-Meijer (✉) · E. Booij · Th. J. Ten Cate
Center for Research and Development of Education, University Medical Center Utrecht,
P.O. Box 85500, 3508 GA Utrecht, The Netherlands
e-mail: m.wijnen-meijer@umcutrecht.nl

M. Van der Schaaf
Department of Education, Utrecht University, Utrecht, The Netherlands

S. Harendza
Department of Internal Medicine, University Medical Center Hamburg-Eppendorf,
Hamburg, Germany

C. Boscardin · Th. J. Ten Cate
Department of Medicine, University of California, San Francisco, CA, USA

J. Van Wijngaarden
Department Clinical Skills Training, University Medical Center Utrecht, Utrecht, The Netherlands

readiness for clinical practice of medical graduates. Further exploration of the procedures for entrustment decisions is recommended.

Keywords Argument based approach · Assessment · Authentic simulation · Coping with unfamiliar clinical situations · Entrustment decisions · Readiness for clinical practice · Validity

Introduction

Background

The goal of academic education is to ensure that students acquire knowledge within the domain of their future profession. Besides knowledge, complex skills such as communication skills, organization skills and problem solving skills can be considered essential for professionals (Birenbaum and Dochy 1996; Fraser and Greenhalgh 2001). In medical education there is growing support for the idea that trainees should be able to apply knowledge to solve clinical problems (Kreiter and Bergus 2008; Ten Cate et al. 2010; Wittert and Nelson 2009).

While standardized assessment of factual knowledge and practical skills is common, medical educators have increasingly realized that such tests have their limitations. For instance, problems are reported regarding the translation of test outcomes to the real world of health care (Arnold 2002; Epstein 2007; Ginsburg et al. 2010; Ginsburg 2011; Howley 2004; Newble 2004; Wass et al. 2001). This has led to a call for a higher validity of assessment procedures in clinical education (Crossley et al. 2011; Govaerts et al. 2007; Tavares and Eva 2012; Wetzel 2012).

"Miller's pyramid" is a well known metaphor that delineates levels of assessment in medical training (Miller 1990). Miller's third level ("shows how") and level 4 ("does") reflect work behavior in a standardized setting and in the actual workplace, respectively. The Objective Structured Clinical Examination (OSCE), in which trainees address structured case-fragments sequentially (Harden and Gleeson 1979), is an example of an assessment at level 3. The strength of the OSCE is its potential to provide evidence of reliability due to its high degree of standardization. Assessment methods at level 4 aim to capture what a trainee actually does in clinical practice. Examples are the mini-clinical evaluation exercise (mini CEX, Norcini et al. 1995) and Direct Observation of Procedural Skills (DOPS, Barton et al. 2012). These kinds of assessment provide useful information about trainees' readiness for clinical practice. Because the clinical context cannot be standardized, precludes high levels of assessment reliability.

Faced with the complexity of a broad, valid assessment of graduates in the real clinical workplace, with the limitations of fragmentation limited authenticity of a simulation, such as an OSCE, we decided to design a semi-standardized assessment procedure. On the one hand, the assessment included as many authentic ingredients of the workplace as possible, on the other hand we used standardized patient scenarios. In Miller's terms, we consider this a *near-does* level.

In this study, we deliberately had not in mind to build a practical assessment tool, but to first include everything necessary to mimic the real world, however, with standardized scenarios to enable the comparison between graduates. We attempted to develop a simulation that reflects the authentic clinical situation better than any other test we know of. Critical in our approach was that we acknowledged the ever-changing, unpredictable

clinical context. Coping with real life clinical problems means coping with unfamiliar situations. To summarize, we wanted to assess whether medical trainees are ready to be entrusted with unfamiliar clinical tasks and called the assessment the "Utrecht Hamburg Trainee Responsibility for Unfamiliar Situations Test" (UHTRUST).

Entrustment is an emerging concept in the thinking of assessment in postgraduate medical education, notable because of the introduction of Entrustable Professional Activities (EPAs), units of professional practice to be entrusted to trainees once they demonstrate the required competence (Ten Cate 2005; Ten Cate and Scheele 2007; Boyce et al. 2011; Chang et al. 2012). This study was not developed to be a practical tool for every day assessment of graduates but rather as a full procedure with a different approach to assessment in the simulated workplace and it is the first we know of that explicitly uses the EPA concept in assessment.

The aim of this paper is to evaluate its validity by building a validity argument for this UHTRUST assessment procedure.

Besides reliability and validity as psychometric aspects of assessments, several authors also mentioned the importance of utility aspects, such as costs, feasibility and acceptability (Van der Vleuten 1996; Messick 1995). We will also pay attention to these descriptors in this study.

The UHTRUST assessment procedure

We developed an authentic simulation for medical graduates at MD level to evaluate their readiness for clinical practice. Our key question was: are medical graduates ready to be entrusted with critical clinical activities (EPAs).This matches the idea that one important characteristic of medical professionals is that they must be trusted to work without supervision (Freidson 1970).

During the assessment, candidates were situated in the role of a beginning resident on a very busy day with the initial instruction: "This is your first day as a resident, at a ward which is yet unknown to you. Unfortunately, your supervisor is called away. It is not possible to cancel the patient appointments, so you will be responsible for them, but you can call your supervisor for help whenever you feel the need to."

The assessment procedure consisted of three phases. In the first phase (1 h), the candidates saw five consecutive standardized patients that "had been admitted to the hospital" with non-common medical problems. During the second phase (3 h) the candidates had time to gather information from the internet, to request lab results and X-rays, to determine differential diagnoses and to draw up a management plan to enable a presentation of each patient to the supervisor at the end of the day. During this phase the candidates also faced distracting tasks, such as reported changes in a patient's condition, questions from nurses or junior students and an urgent organizational problem that needed to be solved. The candidates had the opportunity to call their supervisors by phone if needed and halfway the day there was a brief face-to-face meeting with the supervisor to discuss the candidate's progress. In the third phase (30 min), the candidates reported their differential diagnoses and management options. Each candidate was independently assessed by two or three clinicians, one nurse and six standardized patients (SPs—five of them acted as patients and the sixth played the role of the husband of one of the patients) on different facets of their competence, resulted from a Delphi study among clinical educators (Wijnen-Meijer et al., accepted for publication). One of the clinicians acted as the candidate's personal supervisor during the assessment. The second clinician was present all day and listened into telephone conversations and face-to-face conversations between the supervisor and the candidate.

The third clinician only observed the candidate during the reporting phase. The nurses observed the candidates during the second phase and they deliberately disturbed the candidates in a systematic way with several distracting, but real life tasks.

After all observations the clinicians were asked to individually indicate how much supervision they estimated this trainee would require on nine EPAs, unrelated to the observed scenarios. Table 1 provides a schematic overview of UHTRUST.

We selected the patient cases and distracting tasks in such way that they covered the breadth of the medical profession to a large extent, while at the same time the assessment fitted within 1 day. Similar to an OSCE, the tasks were standardized as much as possible. Differences were the addition of distracting tasks, the fact that the different cases ran simultaneously and the possibility to interact with a supervisor, which provides more similarity with clinical practice.

Kane's quality requirements regarding assessments

High stakes assessments need to be reliable and valid. Reliability refers to consistency of scores across repeated measurements and is considered an important condition for validity. Hence reliability is a component of the validity argument (Wass and Archer 2011; Holmboe and Hawkins 2008). This definition of reliability has remained generally stable during the past decades (Brennan 2006). In contrast, the concept of validity has changed over the years. Brennan (2006) describes how the concept was first defined in terms of the predictive power for future performance (Cureton 1951). In the nineteen seventies and eighties, Cronbach (1971) and Messick (1989) introduced the "unitary" notion of validity. According to their view, validity focuses on the appropriateness of *inferences* from test scores, and not just on the validity of the assessment instrument itself. According to the Standards for Educational and Psychological Measurement (Brennan 2006, p. 2) the preferred definition of test validity is "the degree of which evidence and theory support the interpretations of test scores entailed by purposed used of tests".

Kane (1992) introduced the "argument-based approach" to validity, which is consonant with Messick's ideas of validity, but focuses on the methodology for validation. The goal is to provide a structured and coherent analysis of all evidence in support of the interpretation of test scores. The argument that underpins validity leads from the test administration to

Table 1 Schematic overview of UHTRUST

		Phase 1		Phase 2		Phase 3	
Activities	Briefing	Short meeting with supervisor Consultation five patients	Walking to next location	Collection of diagnostic information about five patients Seven distracting tasks Halfway meeting with supervisor If needed: calls with supervisor Drawing up management plans	Walking to next location	Report and discuss examination- and treatment plans	Debriefing
Duration	30 min	1 h	10 min	3 h	10 min	30 min	30 min
Assessors		Standardized patients		Nurse Clinicians 1 and 2		Clinicians 1, 2 and 3	

the interpretation of scores. Kane (1992, 2006) labels four components in this inferential chain: scoring, generalization, extrapolation and interpretation. The *scoring* component of the argument requires evidence regarding the appropriateness of the assessment conditions, data collection and scoring procedure. The *generalization* component requires evidence that the observed score, coming from the task sample used, is generalizable to a broader domain i.e. the test domain. In the *extrapolation* component, the interpretation is extended to the practice domain. This requires evidence of the link between the data collected in the test and the behaviors of interest in the real world. The fourth component of the argument concerns *interpretation*. Here, a final conclusion is drawn: what implications logically result from the observed skill level of the candidate?

The aim of study reported in this paper is to provide a validity argument for the UHTRUST-assessment procedure, following an argument-based approach for validation. We have chosen this broader approach to integrative interpretations based on different types of relevant evidence because of its appropriateness for complex assessment methods like UHTRUST. We supplemented the argument-based approach for validation with our findings regarding utility aspects.

Method

Participants

Four educationalists developed the assessment procedure together with two recently graduated doctors and six experienced physicians. During the assessment days at two locations in July and August of 2011, 60 candidates participated (30 from the Netherlands and 30 from Germany). All candidates had just completed the medical school program at the moment of the assessments. They participated voluntarily and had applied in reaction to announcements. In Utrecht (the Netherlands) the candidates were assessed by in total 14 physicians, four nurses and six SPs. In Hamburg (Germany) 11 physicians, four nurses and 18 SPs were involved. All candidates were assessed by at least two assessors; twenty Dutch and six German candidates were also judged by a third clinician. The difference between the numbers of SPs is caused by the fact that in Utrecht all 30 times the six patient roles were consistently played by the same SPs, while in Hamburg every role rotated between three SPs. The physicians and nurses were invited to participate, based on their clinical experience and experience with supervising trainees. Furthermore, at each location about 30 persons assisted with the organization of the assessment.

Scoring instruments

The physicians completed three kinds of scoring forms for each candidate. One scoring form included seven so called "facets of competence" (FOCs) that can be considered key in making entrustment decisions by supervisors about residents (see Table 2). These facets were developed during a Delphi and ranking study among physician supervisors in the Netherlands and Germany (Wijnen-Meijer et al., accepted for publication; Wijnen-Meijer et al., submitted). For each FOC, the candidates were scored on a 3-point Likert scale of 1 (weak) to 3 (good) for each of five different patient cases. Additionally, the assessors gave an overall score for each FOC on a 5-point Likert scale, from 1 (very weak) to 5 (very good). The second questionnaire consisted of nine so called "Entrustable Professional Activities" (EPAs; see Table 3), tasks that are suitable to entrust to a trainee once

Table 2 Quality measures FOCs

FOC	Reliability (Phi-coefficient)		Percentage of variance accounted by candidate		Percentage of variance accounted by assessor	
Number of assessors (clinicians)	2	3	2	3	2	3
1. Scientific and empirical grounded method of working	0.88	0.88	71	70	3	4
2. Knowing and maintaining own personal bounds and possibilities	0.85	0.84	65	64	7	5
3. Teamwork and collegiality	0.79	0.73	56	47	13	11
4. Verbal communication with colleagues and supervisors	0.85	0.80	64	57	9	10
5. Responsibility	0.88	0.87	70	69	1	0
6. Safety and risk management	0.74	0.74	48	48	10	9
7. Active professional development	0.89	0.88	72	71	0	0

sufficient capability is attained for unsupervised practice (Ten Cate 2005). The physicians were asked to indicate on a 5-point scale how much supervision they estimated the candidate needs for these EPAs. (1 = he/she is not able to do this; 2 = he/she is able to do this under direct supervision; 3 = he/she is able to do this if supervision is available; 4 = he/she is able to do this independently; 5 = he/she is able to supervise others in performing this activity). The EPAs did not reflect the activities or pathologies involved in the actually observed activities. The third form was a so-called "Post Patient Encounter Form" (PPEF), based on the Post-Encounter Form designed and validated by Durning et al. (2012). The candidates summarised on this form for each patient case the most important problems, differential diagnoses and a proposal for treatment. The assessing physicians scored these aspects on a 5-point Likert scale from 1 (below expectation) to 5 (above expectation). The fourth scoring form was completed by the nurses. This scoring form contained six FOCs, which were similar to six of the seven FOCs that were scored by the clinicians. For each FOC, the candidate was scored by a nurse on a 3-point Likert scale of 1 (weak) to 3 (good) for their performance regarding five different disturbances. Additionally, the nurse gave an overall score for each FOC on a 5-point Likert scale, from 1 (very weak) to 5 (very good). The SPs completed the so called CARE-questionnaire, a validated instrument consisting of 10 questions to measure consultation skills and empathy (Mercer et al. 2004). The SPs scored the candidates on these items with a 5-point scale from 1 (poor) to 5 (excellent).

Evaluation forms

Three months before the assessment days were scheduled, we organized pilot assessments in Utrecht and in Hamburg. At the end of these pilots, all candidates, clinicians, nurses, SPs and staff members evaluated the organization and content of the assessment. We used this information to make adjustments to the assessment and to gather information for the argument for validity.

Procedure to the development of the argument for validity

We developed a theoretical framework for the validity argument based on various theoretical and empirical studies related to the argument-based approach to validity

Table 3 Quality measures "Entrustable professional activities"

EPA	Reliability (Phi-coefficient)		Percentage of variance accounted for by candidate		Percentage of variance accounted for by assessor	
Number of assessors (clinicians)	2	3	2	3	2	3
1. Emergency assistance with acute cardiac failure	0.71	0.70	45	44	9	7
2. Handling a patient complaint	0.48	0.42	23	19	44	40
3. Pre-operative information and consent	0.50	0.49	25	24	37	36
4. Breaking bad news	0.27	0.36	11	16	56	56
5. Clinical reasoning under time pressure	0.71	0.68	45	42	17	14
6. Solving a management problem	0.62	0.63	36	36	36	35
7. Suspicion of self-induced disease	0.51	0.48	26	23	28	29
8. Handling of a seriously ill patient	0.63	0.61	36	34	40	41
9. Interaction with a consultant	0.71	0.72	45	46	13	12

(Bakker 2008; Chapelle et al. 2010; Hawkins et al. 2009; Holmboe and Hawkins 2008; Kane 2004, 2006). The framework illustrates the four major inferences that are associated with an argument-based approach to validity and their underlying assumptions (see Table 4).

Table 4 Framework for an argument for validity—inferences, assumptions and warrants licensing the assumptions

Inferences and assumptions		Warrants licensing the assumptions
Inference 1	*Scoring: from the observed performance to the observed score*	
Assumption 1.1	The assessment conditions are appropriate	The assessment conditions were standardized, so candidates were provided with equal opportunities to show their abilities and test scores could be compared to one another A detailed planning was defined to ensure a smooth running of the assessment day. This planning was evaluated after the pilots It was unlikely that the candidates had access to the assessment tasks prior to the assessment days
Assumption 1.2	The scores are recorded accurately	All candidates were judged by multiple trained assessors. The assessors judged the candidates' behavior on seven FOCs and nine EPAs Assessors were urged to follow a systematic and transparent scoring procedure. This reduces the risk of invalid and unreliable judgments Instructed staff members were assigned to the different groups of assessors. These staff members checked whether the scoring forms were filled out correctly and answered assessors' questions In every room staff members were present to prevent the candidates to exchange information regarding the cases Security measures were taken to prevent loss of data and to protect the assessments' integrity

Table 4 continued

Inferences and assumptions		Warrants licensing the assumptions
Assumption 1.3	The scoring criteria are appropriate and acceptable	A panel of informed experts agreed upon the content and language of seven FOCs. These were used to develop an analytic scoring rubric Global rating scales were used to score the FOCs. They were expected to be more feasible, equally reliable, and more valid than detailed (dichotomous) behavioral checklists Experts developed nine EPAs that were used to indicate to what extent the candidates can be entrusted with critical clinical tasks
Assumption 1.4	Reliable and valid scoring of the performance by the assessors	The sample and selection procedure of the assessors was acceptable During a "frame of reference" training assessors attempted to reach shared understanding of the content and performance standards. Assessors were also informed on how to avoid typical rater errors The internal consistency (reliability), calculated by means of phi-coefficient, of the raters for the FOCs varies from acceptable to good (see Table 2). The reliability of the raters for the EPAs varies from moderate to acceptable (see Table 3)
Inference 2	*Generalization: from the observed score to the expected universe score*	
Assumption 2.1	The scores are stable and random error due to different occasions, raters and tasks is controlled	The standardization measures described in inference 1 controlled the random error caused by administration occasion, rater and tasks Multiple assessors per candidate were used to reduce the influence of personal biases of the individual assessors Candidates were confronted with multiple cases. This reduced the variance caused by tasks specificity, and provided the candidates with the opportunity to demonstrate their competences on multiple occasions G-studies are conducted to determine the percentage of variance that can be explained by the candidate and by the assessor. For the FOCs the percentage of variance accounted by the candidate is relatively high (see Table 2); for the EPAs this percentage is lower (see Table 3)
Assumption 2.2	The sample of observations is representative of the universe of generalization	It was only possible to sample a relatively small number of assessment tasks. In order to compensate this deficiency, serious effort was made to draw a representative sample from the universe of generalization Experts were consulted during the task development and evaluation. They made a blueprint for the content of the assessment in order to make sure that the task sample could not be completed without the use of the defined FOCs

Table 4 continued

Inferences and assumptions		Warrants licensing the assumptions
Inference 3	*Extrapolation: from the universe score to the expected level of skill in the target domain*	
Assumption 3.1	The universe score is related to the level of skill of the graduate in the target domain	The authentic character of the assessment makes the argument for extrapolation plausible When a comprehensive construct is measured, the practical limits of assessment must be accepted UHTRUST provided negative evidence on the candidates' true ability to cope with unfamiliar clinical situations
Assumption 3.2	There are no systematic errors that are likely to undermine the extrapolation	The standardized assessment conditions and the use of standardized patients brought about an artificial aspect to the assessment Sources of irrelevant variance caused by systematic differences between SPs (and real patients) and time pressure were identified and controlled
Inference 4	*Interpretation: from the level of skill in the target domain to the test interpretation*	
Assumption 4.1	All assumptions are defensible with accurate and plausible evidence	Most assumptions were defendable with accurate (and often parallel lines of) backing
Assumption 4.2	The data acquired by the assessment can be used for the intended purposes	Because most validity assumptions were defendable with accurate and often parallel lines of backing, this assessment can be used for the intended purpose

The writing of the argument for validity was an iterative process. All pieces of validity evidence were collected and arranged in a way that did justice to Kane's argument-based approach to validity. During the development of the assessment, we discussed and wrote down all considerations and decisions. We used all available data to design a coherent and plausible argument. To do so, the data were linked to one or more of the four major inferences in the argument for validity.

Analysis

We analyzed the results from the scoring forms by means of Pearson correlation coefficients between all FOCs and EPAs. In addition, we conducted a generalizability analysis on the FOCs and EPAs to determine the proportion of variance accounted by the candidates and by the assessors. We also calculated means and standard deviations of the scores on the evaluation forms.

Ethical approval

Ethical approval for the Dutch part of the study was obtained from the NVMO Ethical Review Board. For the German part, ethical approval was obtained from the State of Hamburg Physicians' Ethics Board.

Results: argument for validation of UHTRUST

We structured the results section to address the four inferences and ten assumptions of Kane's validity argument. A summary of the inferences is presented in Table 4.

Inference 1: Scoring-from the observed performance to the observed score

Assumption 1.1: The assessment conditions are appropriate

We implemented the UHTRUST assessment procedure on two different locations and with two different groups of candidates, physicians, nurses and standardized patients (SPs). To enhance similarity between the two administration occasions, the observations of the assessment were made under semi-standardized assessment conditions (Cohen and Wollack 2006; Kane 2006). Before the assessment days test developers prepared standardized instructions (e.g., time limits), conditions of administration and guidelines for scoring. This implied how a valid response or judgment had to be constructed, what ancillary materials were allowed for the candidates and how much help supervisors and nurses were expected to provide. We ensured during the assessment days that all candidates were assessed with the same clinical content and tasks and under the same conditions.

UHTRUST was meant to be a realistic assessment with open-ended tasks. In this kind of assessment it is difficult to discern all potential threats to standardization, "including those associated with SP portrayal, unanticipated student reactions to the scripted SP responses and case irregularities" (Holmboe and Hawkins 2008, p. 105). To enhance standardization, we used the evaluations and recordings made during the pilot assessments to optimize the cases and strengthen the effectiveness of instructions for all participants.

To ensure smoothly running assessment procedures we constructed a detailed planning. This planning was partly based on empirical and theoretical studies containing useful advice derived from successfully run authentic assessments (Boursicot and Roberts 2005; Cohen and Wollack 2006; Holmboe and Hawkins 2008). Besides detailed time schedules for candidates, physicians, nurses, SPs and staff members, it included descriptions of necessary practical and logistical arrangements, to begin with the pilot study on two sites. Its evaluations among both Dutch and German participants ($N = 84$) revealed that the pilot days were felt well organized. Combined mean score on this item was 4.11 ($SD = 0.44$) on a 5-point scale.

We made sure candidates had no access to the assessment tasks prior to the assessment administration, as the scores of these candidates would then not accurately reflect their ability levels. We developed the cases of UHTRUST developed especially for this assessment in the months prior to the assessment days. Premature exposure to the assessment tasks was therefore unlikely. One case about hemolytic uremic syndrome was replaced as an unusually large outbreak of this disease at one site (Hamburg) took place between the pilot study and the main study (Harendza 2011).

Assumption 1.2: The scores are recorded accurately

The use of a systematic and transparent scoring procedure reduces the risk of invalid and unreliable judgments (Bakker 2008; Kane 2006). Prior to the assessment, all assessors received a frame of reference training (Holmboe and Hawkins 2008), including explanations about the impact of scoring errors and biases. In addition, all assessors were orally

trained in using the scoring forms and received written instructions about the scoring procedures. At both assessment days, instructed staff members were assigned to the different groups of assessors. They checked whether the scoring forms were filled out correctly and readable, and were available to answer any assessors' questions.

We took security measures to prevent loss or mixing up of data. At the beginning of the day, all candidates, assessors, SPs and staff members checked-in. At the end of the day all participants checked-out and staff members made certain that all intended documents had been received.

UHTRUST was administered for research purposes only. Since there was no impact on academic progress or graduation or any risk of harmful consequences for those with high or low scores, and because participation was voluntary, it was assumed that the candidates' motivation to cheat was low. To be sure, in every room staff members were present to prevent candidates to exchange information regarding cases.

Assumption 1.3: The scoring criteria are appropriate and acceptable

Scoring criteria (i.e., content standards) indicate what candidates should know or be able to do. They guide the assessors' judgment about the quality of a candidates' performance (Gipps 1994). The development of the facets of competence (FOCs) was split into two phases. First, in a Delphi study ten FOCs were identified as most important to entrust critical clinical tasks to a trainee (Wijnen-Meijer et al., accepted for publication; Wijnen-Meijer et al., submitted). This study was conducted in the Netherlands and afterwards validated in Germany. Second, in the main study seven FOCs were scored by the physicians and nurses and three FOCs were judged by the SPs. The FOCs were not further specified in sub-criteria, resulting in rather global assessment criteria, which are commonly used for the measurement of discrete constructs (e.g., communication, empathy) as well as in assessing more broad constructs (e.g., the ability to take responsibility) (Hawkins et al. 2009). Detailed (dichotomous) checklists (e.g., makes eye contact, introduces themselves) for such broad constructs often fail to validly capture essential features, due to a difficulty to quantify elements of expert behavior (Holmboe and Hawkins 2008).

At the end of the day, the assessing physicians were asked to indicate for each observed candidate to what extent they would entrust this person with new critical clinical activities. In order to do so, a variety of nine EPAs was identified (see Table 3). The developers made sure that each of the identified FOCs would be necessary in at least one EPA, to ensure that all EPAs together covered all FOCs. Pearson correlations were calculated to verify these relationships (Table 5). All FOCs correlated significantly with all EPAs ($p < .01$).

Assumption 1.4: Reliable and valid scoring of the performance by the assessors

For performance assessments, the quality of the assessment as a whole is related to the ability of the assessors to use the scoring criteria to reach a technically and professionally defensible conclusion (Dwyer 1995). We selected all physicians and nurses involved in UHTRUST because of their active clinical- and supervising experience, to ensure they are capable to make profound judgments about a candidate's ability. All assessors participated voluntarily. This is important, because high motivation contributes to the quality of the rating outcomes (Govaerts et al. 2007).

To prepare the assessors for judging and scoring the performance of the candidates, a frame of reference training was delivered (Holmboe and Hawkins 2008). The training was standardized and was given by the same two instructors. The training sessions at both

Table 5 Pearson correlation coefficients EPAs–FOCs

	FOC 1	FOC 2	FOC 3	FOC 4	FOC 5	FOC 6	FOC 7
EPA 1	.686	.605	.471	.552	.570	.581	.545
EPA 2	.476	.363	.418	.407	.458	.466	.375
EPA 3	.502	.507	.423	.393	.495	.497	.429
EPA 4	.407	.317	.362	.311	.389	.369	.262
EPA 5	.628	.587	.445	.537	.605	.615	.540
EPA 6	.553	.512	.391	.451	.530	.458	.485
EPA 7	.507	.464	.408	.379	.486	.485	.471
EPA 8	.564	.494	.427	.441	.516	.448	.524
EPA 9	.558	.493	.465	.516	.561	.502	.455

All correlations are significant at the 0.01 level (2-tailed)

locations were similar. During the training, the assessors were taught to apply the detailed scoring procedure in a systematic and consistent way. The assessors had to elaborate and share conceptualizations of what constitutes competent behavior and were asked to formulate standards for acceptable and unacceptable behaviors for each of the FOCs. This was important given the fact that the FOCs refer to abstract qualities which assessors need to infer from the performance of the candidate. The assessors reached consensus about performance standards and rating scale anchors for each of the scoring criteria. According to Knight (2002), there will always be a certain degree of ambiguity about the meaning of criteria and interpretation of the standards. However, it is important to put trust in the judgment of an expert and not to quell creativity (Gipps 1994; Ten Cate 2005). The scoring criteria, the detailed scoring procedure and the assessor selection procedure and training were all designed and implemented to maximize objective and reliable scoring. However, they do not yet guarantee a high quality assessment processes (Nijveldt 2007). That is why the effects of these measures where statistically examined.

To estimate the reliability of the clinician raters on the seven FOCs, the EPAs and the PPEFs we conducted a G-study, for both two physicians (who observed the candidate all day) and three physicians (including the assessor who only observed the reporting phase). The reliability (phi-coefficient) of three raters for the FOCs varies from acceptable (.73) to good (.88) (see Table 2). The reliability of three raters for the EPAs varied from moderate (.36) to acceptable (.72) (see Table 3). A G-study was also conducted for the PPEFs that summarised for each patient case the most important problems, differential diagnoses and a management proposal. The reliability (phi-coefficient) for both two and three physicians varied from acceptable (.63 and .64, respectively) to good (.90 and .89, respectively) except for "Problem" of case 3 (.39 and .42, respectively) and "DD" of case 5 (.59 and .48, respectively) (see Table 6). For all scores there is little difference in reliability when two or three assessors are included).

Inference 2: Generalization-from the observed score to the expected universe score

Assumption 2.1: The scores are stable and random error due to different occasions, raters and tasks is controlled

Any facet that is allowed to vary in the universe of generalization (e.g., tasks, assessors) and that is sampled by the measurement procedure contributes to random error of an

assessment score (Kane 2006; Lane and Stone 2006). We implemented the standardization measures, as described in the section about the scoring inference, to reduce for random error caused by three variables: the administration occasion, the assessors and the tasks. Furthermore, we took measures to increase stability of scoring. First, the stability of scoring was enhanced by the use of multiple assessors per candidate. This reduces the influence of personal biases of individual assessors (Kane 2006). Second, the use of multiple cases compensates psychometric limitations inherent to a single case, assuming that all cases have adequate quality. The candidates encountered five patient cases and seven distracting tasks, designed to sample skills broadly over the course of the assessment day. The use of multiple cases also enhanced the accuracy of the test scores, because the assessors were given the opportunity to judge the candidates based on their performance on multiple occasions, and the candidates were provided with several opportunities to demonstrate their competence.

In a G-study, we calculated the overall variance of the test for multiple sources. If three assessors are included, we find percentages of variance explained by the candidate to be relatively high for the FOCs (varying from 47 to 71 %, see Table 2). For the EPAs they were lower (16 to 46 %, see Table 3) and for the PPEFs (Table 6) there is more divergence, particularly for the items "problems" (19–66 %) and "DD" (24–74 %);for "treatment" they were more consistent and relatively high (41–67 %). In addition we found that percentages hardly differ if either two or three assessors are included).

Table 6 Quality measures "Post-Patient Encounter Forms": problems, DD and treatment per patient case

	Reliability (Phi-coefficient)		Percentage of variance accounted for by candidate		Percentage of variance accounted for by assessor	
Number of assessors (clinicians)	2	3	2	3	2	3
Description of problems						
Case 1	0.63	0.64	37	37	28	23
Case 2	0.84	0.86	63	66	15	10
Case 3	0.39	0.42	18	19	28	23
Case 4	0.73	0.75	48	50	20	16
Case 5	0.68	0.66	42	39	2	8
Differential diagnosis						
Case 1	0.85	0.77	66	53	8	15
Case 2	0.85	0.86	65	68	9	9
Case 3	0.67	0.75	40	50	17	6
Case 4	0.90	0.89	76	74	4	2
Case 5	0.59	0.48	32	24	0	0
Proposal for treatment						
Case 1	0.85	0.79	65	55	0	13
Case 2	0.78	0.68	54	41	5	9
Case 3	0.84	0.86	63	67	0	0
Case 4	0.84	0.80	64	57	19	22
Case 5	0.71	0.71	45	45	0	0

Assumption 2.2: The sample of observations is representative of the universe of generalization

According to Bakker (2008), the selection of representative samples of assessment tasks is an important issue in performance assessments. To make sure that the UHTRUST tasks would cover the selected FOCs and to make a blueprint for the content of the assessment, the test developers consulted ten medical experts. The experts agreed that all tasks included in UHTRUST together portrayed a sufficiently broad content. They also advised on how cases should play out and.

The raters' judgments about the candidates' ability to take responsibility for unfamiliar situations were also thought to be influenced by the way candidates would handle the disruptions during the second phase of the assessment. The experts and test developers also thoroughly planned and discussed these additional tasks.

Inference 3: Extrapolation-from the universe score to the expected level of skill in the target domain

Assumption 3.1: The universe score is related to the level of skill of the graduate in the target domain

We made serious efforts to achieve a high level of physical and psychological fidelity. First, the choice to work with SPs instead of written patient scenarios, contributed to the level of realism. In performing the simulation, the SP does not only present the gestalt and history of the patient being simulated, "but the body language, physical findings and emotional and personality characteristics as well" (Cleland et al. 2009, p. 478). The psychological fidelity was further enhanced by the tasks and the modes of presentation. Lane and Stone (2006) stated that these kinds of high-fidelity tasks can easily be translated to expected performance in the real world. The tasks were designed in such a way that they could also occur on a real clinical ward, including referral letters and the opportunity to request for lab- and radiology results. Finally, the act of observing can interfere with the level of authenticity. However, the candidates were never observed by assessors at unrealistic moments. For example, none of the assessors were present during the patient encounters. The effect of our efforts to maximize authenticity was evaluated after the pilot assessments. On the evaluation forms both candidates and assessors were asked whether the pilot had a high level of authenticity. The candidates' ($N = 18$) mean score on this item was 4.3 ($SD = .66$) and the assessors' ($N = 20$) mean score was 3.9 ($SD = .55$) on a 5-point scale. Another question was whether the assessed competences were relevant for clinical practice. The mean score of the candidates ($N = 20$) on this item was 4.4 ($SD = .59$) and of the assessors ($N = 19$) 4.15 ($SD = 1.3$) on a 5-point scale.

The basic assumption is that all activities in the test domain are necessary for the effective dealing with unfamiliar clinical situations in the practice domain. This assumption makes it reasonable to expect that candidates who were successful in the assessment would also be successful in unfamiliar situations in reality. However, this cannot be taken as absolute evidence. When a candidate demonstrates that he or she is capable to take responsibility for unfamiliar situations during UHTRUST, this is no guarantee that this behavior will be manifested in real clinical settings, e.g., when a candidate lacks other skills or features (such as motor coordination) that were not included in the test. Such limitations of the validity of assessments must be accepted (Knight 2002). Kane (2004) states that for most performance assessments this assumption tends to be stronger on the

negative side. Even though not every aspect of the construct can be measured, it is reasonable to assume that a candidate who showed serious deficiencies in the test domain would also show deficiencies in the practice domain.

Assumption 3.2: There are no systematic errors that are likely to undermine extrapolation

As mentioned above, the assessment and scoring conditions were fixed for all candidates. Some of these standardization measures brought about an artificial aspect to UHTRUST and resulted in sources of systematic error that had to be identified and controlled.

First, also trained SPs can be a source of construct irrelevant variance. Holmboe and Hawkins (2008) stated that even though little research on the subject exists, it is inevitable that differences between SPs and real patients occur. For this reason we checked the acceptability of the performance of the SPs during the pilot assessments on the evaluation form. Candidates (N = 20) gave the plausibility of the SPs' performances a mean score of 4.55 (SD = .58) on a 5-point scale. Furthermore, based on the recordings of the encounters in the pilot study, we improved the training for the SPs.

Second, time pressure can yield invalid measures of proficiency, contributing to construct-irrelevant variance (Holmboe and Hawkins 2008). The medical experts that were involved in the development process indicated that the time allotted for the candidates was short, but realistic. On the evaluation form, candidates of the pilot (N = 20) were asked if they had had enough time available to complete the tasks in the individual phases of the assessment. On average, the candidates were quite satisfied with the amount of time they had for various phases. For all phases, the mean score was higher than 3 on a 5-point scale. During the evaluation of the pilot assessments, various candidates and assessors indicated that time pressure is often present in clinical practice, and therefore should not be seen as irrelevant variance.

Inference 4: Interpretation-from the level of skill in the target domain to the test domain

Assumption 4.1: All assumptions are defensible with accurate and plausible evidence

To make sure that the most crucial validity assumptions were critically considered and substantiated with accurate and plausible backing, we made the inferences and assumptions that underlie an argument for validity explicit in a theoretical framework in advance (Table 4). The aim of this study was to evaluate the most prominent aspects of a performance assessment: planning, standardization measures, scoring procedure, reliability of the scores, the authenticity level and the investigation of potential threats to validity (Holmboe and Hawkins 2008).

Assumption 4.2: Data acquired by the assessment can be used for the intended purposes

We developed UHTRUST to answer the question to what extent medical graduates can be entrusted with clinical tasks in unfamiliar situations. As most validity assumptions were defendable with accurate and often parallel lines of backing, UHTRUST can be used for the intended purpose: the formative assessment of the readiness for clinical practice of medical graduates.

Utility aspects

For educational practice, also utility aspects are important. The acceptability of UHTRUST among the participants was good, as all persons involved were positive about it. This can be derived from the results of the evaluation form. Concerning the feasibility and costs of UHTRUST can be said that implementation requires considerable investments in regard to time and effort of staff and assessors and, as a result, considerable finances. These are comparable with the investments needed for the implementation of an OSCE (Boursicot and Roberts 2005). Whether this is too much is a difficult to answer question, as it depends on the interest one has in a valid outcome. Studies to establish predictive validity can help to support such decisions.

Discussion

UHTRUST is an authentic assessment procedure that intends to measure a broad and complex construct i.e. the extent to which medical graduates can be entrusted with (unfamiliar) critical clinical activities. In the current study Kane's argument-based approach to validity was used to write an argument to support the validity of UHTRUST. The construction of an argument for validity is an iterative process that should lead to continued improvement in the quality and defensibility of an assessment (Kane 2006). We found that most validity assumptions were defendable with accurate and often parallel lines of backing. Based on this argument we conclude that the assessment can be used for the intended purpose.

One of the weaker components of the validity argument of UHTRUST is the reliability of the EPA scores, which is moderate for some EPAs. This is the most novel part of the assessment procedure, as we not only asked to evaluate observed behavior, but also to determine trust in future behavior. We are not aware of other procedures that attempt to measure this construct. It can be assumed that uncertainty among assessors increases if they must make inferences to predict unobserved behavior. Sterkenburg et al. (2010) found that anesthesiology physicians value substantial acquaintance with trainees highly as a condition to trust them. In addition, context variables, such as time of the day, hospital personnel and facilities available influence entrustment decisions (Dijksterhuis et al. 2009; Sterkenburg et al. 2010). This aspect needs attention in future applications of UHTRUST. Further improvement of the training or instructions for assessors and additional information about the candidates could increase the reliability of the EPA scores, as well as including context limitations.

To determine the utility of UHTRUST, the 'utility equation' of assessments (Van der Vleuten 1996) is useful. As stated above, the reliability and validity evidence is acceptable. Furthermore, the educational impact is good: candidates noted that the assessment gave them a lot of information about their strengths and weaknesses. They considered UHTRUST the "most comprehensive examination ever encountered" (quote by one candidate). Also the acceptability appears to be good: the assessors and candidates were very positive about this assessment. Despite the considerable investments regarding time and finances that are needed for the implementation of UHTRUST, which are comparable with an OSCE, it is not unfeasible. The results of the G-studies show that the difference in reliability between assessors who observed the candidate all day and those who only observed the candidate during the reporting phase is small. This indicates that it is possible to reduce the required time investments of clinicians. Building on our experiences, a future

step can be to determine which ingredients seem practical to include in the assessment. In addition, as assessors and staff acquire more experience with the assessment, the time investment is likely to decrease.

To provide a validity argument for UHTRUST, we followed Kane's argument-based approach for validation. This approach appeared to be useful to shed a light on the overall validity of complex assessments such as UHTRUST. It makes decisions explicit during the development and implementation of the assessment, and consequently the strengths and weaknesses. It remains disputable whether or not more validity evidence should have been gathered. For example, in our study, interviews with assessors could have given more insight in the quality of their cognitive processes and the effects of the assessor training. More discussion about how much and what kind of validity is needed should therefore be valuable. Kane (2006) stated that it is unlikely that all inferences can be evaluated. The decision which inferences should and which should not be evaluated depends mainly on the purpose of the assessment.

In our opinion, UHTRUST can be implemented for different purposes. One possibility is to compare the readiness for clinical practice of medical graduates from different medical schools, in the context of research or to find out whether curricular change is needed. The assessment can also be used to judge the performance of individual graduates, for instance for the purpose of residency selection for postgraduate training programs. Further exploration of the possible implementations of UHTRUST, including conditions and consequences, is recommended.

References

Arnold, L. (2002). Assessing professional behavior: Yesterday, today and tomorrow. *Academic Medicine, 77*, 502–515.

Bakker, M. (2008). *Design and evaluation of video portfolios. Reliability, generalizability, and validity of an authentic performance assessment for teachers*. Leiden: Mostert & Van Onderen.

Barton, J. R., Corbett, S., & Van der Vleuten, C. P. (2012). The validity and reliability of a direct observation of procedural skills assessment tool: Assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy, 75*(3), 591–597.

Birenbaum, M., & Dochy, F. (Eds). (1996). *Alternatives in assessment of achievement, learning processes and prior knowledge*. Boston: Kluwer.

Boursicot, K., & Roberts, T. (2005). How to set up an OSCE. *The Clinical Teacher, 2*, 16–20.

Boyce, P., Spratt, C., Davies, M., & McEvoy, P. (2011). Using entrustable professional activities to guide curriculum development in psychiatry training. *BMC Medical Education, 11*, 96.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: American Council on Education and Praeger Publishers.

Chang, A., Bowen, J. L., Buranosky, R. A., Frankel, R. M., Gosh, N., Rosenblum, M. J., Thompson, S., & Green, M. L. (2012). Transforming primary care training-patient-centered medical home entrustable professional activities for internal medicine residents. *Journal of General Internal Medicine* (early online).

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement Issues and practice, 29*, 3–13.

Cleland, J. A., Abe, K., & Rethans, J. (2009). The use of simulated patients in medical education: AMEE Guide no. 42. *Medical Teacher, 31*, 477–486.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education, 45*, 560–569.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Dijksterhuis, M. G. K., Teunissen, P. W., Voorhuis, M., Schuwirth, L. W. T., Ten Cate, Th. J., Braat, D. D. M., et al. (2009). Determining competence and progressive independence in postgraduate clinical training. *Medical Education, 43*, 1156–1165.

Durning, S. J., Artino, A., Boulet, J., La Rochelle, J., Van der Vleuten, C., Arze, B., et al. (2012). The feasibility, reliability and validity of a post-encounter form for evaluating clinical reasoning. *Medical Teacher, 34*, 30–37.

Dwyer, C. A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A. J. Shinkfield & D. Stufflebeam (Eds.), *Teacher evaluation guide to effective practice* (pp. 62–80). Boston: Kluwer.

Epstein, R. M. (2007). Assessment in medical education. *The New England journal of medicine, 356*, 387–396.

Fraser, S. W., & Greenhalgh, T. (2001). Coping with complexity: Educating for capability. *BMJ, 323*, 799–803.

Freidson, E. (1970). *Profession of medicine: A study of the sociology of applied knowledge.* New York: Dodd, Mead & Company.

Ginsburg, S. (2011). Respecting the expertise of clinician assessors: construct alignment is one good answer. *Medical Education, 45*, 546–548.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*, 780–786.

Gipps, C. V. (1994). *Beyond testing. Towards a theory of educational assessment*. London: RoutledgeFalmer.

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

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*(1), 41–54.

Harendza, S. (2011). "HUS" diary of a German nephrologist during the current EHEC outbreak in Europe. *Kidney International, 80*, 687–689.

Hawkins, R. E., Katsufrakis, P. J., Holtman, M. C., & Clauser, B. E. (2009). Assessment of medical professionalism: Who, what, when, where, how, and… why? *Medical Teacher, 31*, 348–361.

Holmboe, E. S., & Hawkins, R. E. (Eds.). (2008). *Practical guide to the evaluation of clinical competence*. Philadelphia: Mosby-Elsevier.

Howley, L. D. (2004). Performance assessment in medical education: Where we've been and where we're going. *Evaluation and the Health Professions, 27*, 285–301.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective, 2,* 135–170.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.

Knight, P. T. (2002). The Achilles' heel of quality: The assessment of student learning. *Quality in Higher Education, 8*, 107–115.

Kreiter, C. D., & Bergus, G. (2008). The validity of performance-based measures of clinical reasoning and alternative approaches. *Medical Education, 43*, 320–325.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–432). Westport, CT: American Council on Education and Praeger Publishers.

Mercer, S. W., Maxwell, M., Heaney, D., & Watt, G. C. M. (2004). The consultation and relational empathy (CARE) measure: Development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice, 21*, 699–705.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

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

Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education, 38*, 199–203.

Nijveldt, M. (2007). *Validity in teacher assessment. An exploration of the judgement processes of assessors*. Enschede: Gildeprint.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The Mini-CEX (Clinical Evaluation Exercise): A preliminary investigation. *Annals of Internal Medicine, 123*, 795–799.

Sterkenburg, A., Barach, P., Kalkman, C., Gielen, M., & Ten Cate, O. T. J. (2010). When do supervising physicians decide to entrust residents with unsupervised tasks? *Academic Medicine, 85*, 1408–1417.

Tavares, W., & Eva, K. W. (2012). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education.* doi:10.1007/s10459-012-9370-3.

Ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39*, 1176–1177.

Ten Cate, O. & Scheele, F. (2007). Competence-based postgraduate training: Can we bridge the gap between educational theory and clinical practice? *Academic Medicine, 82*, 542–547.

Ten Cate, O., Snell, L., & Carraccio, C. (2010). Medical competence: The interplay between individual ability and the health care environment. *Medical Teacher, 32*, 669–675.

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

Wass, V., & Archer, J. (2011). Assessing learners. In T. Dornan, K. Mann, A. Scherpbier, & J. Spencer (Eds.), *Medical education: Theory and practice* (pp. 229–255). Toronto: Churchill Livingstone Elsevier.

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet, 357*, 945–949.

Wetzel, A. P. (2012). Analysis methods and validity evidence: A review of instrument development across the medical education continuum. *Academic Medicine, 87*(8), 2012.

Wittert, G. A., & Nelson, A. J. (2009). Medical Education: Revolution, devolution and evolution in curriculum philosophy and design. *Medical Journal of Australia, 191*, 35–37.

Wijnen-Meijer, M., Van der Schaaf, M., Nillesen, K., Harendza, S. & Ten Cate, O. Essential FOCs that enable trust in graduates: A Delphi study among physician educators in the Netherlands. *Journal of Graduate Medical Education* (Accepted for publication).

Wijnen-Meijer, M., Van der Schaaf, M., Nillesen, K., Harendza, S. & Ten Cate, O. Essential facets of competence that enable trust in medical graduates: A ranking study among physician educators in two countries. (Submitted).