

The Utrecht Hamburg Trainee Responsibility for Unfamiliar Situations Test (UHTRUST): the preliminary argument for validation of a clinical performance assessment

STUDENT:

EVELINE BOOIJ, 3439992

FIRST SUPERVISOR:

MARIEKE VAN DER SCHAAF

SECOND SUPERVISOR:

HAVVA JONGEN

COORDINATOR OF THE RESEARCH PROJECT:

MARJO WIJNEN-MEIJER

Utrecht University, Master 'Onderwijskundig ontwerp en advisering'

June 2011

Abstract

The Utrecht Hamburg Trainee Responsibility for Unfamiliar Situations Test (UHTRUST) is a clinical performance assessment that intends to measure a broad and complex construct, i.e. the ability of medical graduates to cope with unfamiliar clinical situations. In this study Kane's *argument-based approach to validity* was used to write an argument for validation for the UHTRUST pilot version. Ten assumptions underlying the four major inferences of the argument-based approach to validity were identified and elaborated. Even though not all data of the pilot assessment were fully processed yet, it was found that most validity assumptions were defensible with accurate, and often parallel lines of backing. For example, the analytic scoring criteria were carefully designed by means of a Delphi study, and a principal component analysis on the analytic scores indicated that these criteria indeed reflected a single construct. However, the pilot assessment also revealed some potential threats to validity. For example, clues were found that some assessors had deviated from the prescribed scoring procedure, and that some of the ratings were biased. These problems should be properly addressed in order to improve the quality and defensibility of the final product. On future registration occasions it is also important to gather validity evidence for the most questionable assumption of every rater-scored assessment, namely the quality of the cognitive processes underlying raters' judgment. Because of limited resources, for the pilot assessment it was not feasible to do so.

1. Introduction

1.1 Problem description and research questions

1.1.1 The vertically integrated curriculum and coping with unfamiliar clinical situations

In the Netherlands medical schools no longer work with traditional curricula, in which theory and practice are strictly separated. Instead, a new vertically integrated (VI) curriculum has been established. The philosophy of this curriculum is that students should be encouraged to develop higher-order thinking skills and should be provided with the opportunity to become self-directed (lifelong) learners. In order to do this, curricular elements are placed in a more clinical context. The VI curriculum encourages early clinical experiences, longer clerkships and increasing levels of responsibility for medical students. This innovation seems to facilitate the transition to post-graduate training, given that graduates of medical schools who have experienced the VI curriculum are found to make definitive career choices earlier and feel more prepared for work and postgraduate training (Wijnen-Meijer, Ten Cate, Van der Schaaf & Borleffs, 2010).

In order to investigate the further effects of the VI curriculum, an authentic performance assessment (PA) is being developed at the University Medical Centre Utrecht (UMCU). An authentic PA is an assessment in which examinees are asked to perform an activity in an authentic context (Gipps, 1994). The purpose of this PA is to assess an examinee's ability to cope with unfamiliar clinical situations. This is an important ability, since continuous and rapid developments in science and technology have made it impossible to provide medical students with all the necessary basic sciences and factual knowledge before graduation (Ten Cate, 2007). Therefore, it is likely that medical graduates are confronted with unfamiliar clinical situations at work.

Given the philosophy of the VI curriculum, it is expected that the VI graduates are better prepared to deal with unfamiliar clinical situations than the non-VI graduates. Since medical schools in the Netherlands all work with the VI curriculum, the non-VI graduates will be medical students of the Universitätsklinikum Hamburg-Eppendorf (UKE), Germany. This is why the assessment is called the 'Utrecht Hamburg Trainee Responsibility for Unfamiliar Situations Test', or shortly the UHTRUST.

1.1.2 The UHTRUST assessment

The UHTRUST is a PA in which content, administration and scoring are standardized, but where attention is paid to high physical and psychological fidelity. In the PA examinees are asked to lead a simulated ward. During their day on this ward they are confronted with five unfamiliar cases. These cases are played by simulation patients (SPs); real-life actors who are trained to portray a patient with a specific condition in a realistic and standardized way (Cleland, Abe & Rethans, 2009). During the information gathering process, the graduates face all kinds of standardized work-related disruptions, such as urgent phone calls, a nurse with questions or a SP in need of emergency care.

At the end of the assessment day the graduates report on their management proposals to their assessors who also observed them during various other occasions on that day. These assessors are medical experts (i.e. experienced physicians and nurses) who are trained to score

the graduates performances according to a detailed scoring procedure. They use their impressions and observations of the graduates to judge their ability to cope with unfamiliar clinical situations on an analytic and holistic scoring rubric. The analytic scoring rubric enables the systematic evaluation of the most critical features in the construct of interest, i.e. 1) scientific and empirical grounded method of working; 2) knowing and maintaining own personal bounds and possibilities; 3) active professional development; 4) teamwork and collegiality; 5) active listening to patients; 6) verbal communication with colleagues and supervisors; 7) empathy and openness; 8) responsibility; 9) coping with mistakes; and 10) safety and risk management. These features were selected by means of a Delphi-study among physicians in Utrecht and Hamburg, and resemble the general exit qualifications for medical students (e.g. The CanMEDS 2005 Physician Competency Framework). The features were used as scoring criteria on the analytic scoring rubric. The holistic evaluations concern entrustment decisions; assessors are asked whether or not they would entrust the graduate to diagnose and treat other unfamiliar cases. The course of the day is standardized so graduates experiences during the assessment do not highly differ; the information and time available, the patient scenarios, the disruptions, and the mode of presentation are all determined beforehand as much as possible.

1.1.3 The validity of the UHTRUST pilot version

PAs can provide a rich view of students' competences in context, and are often seen as a highly valid assessment instrument. However, validity cannot be taken for granted. Messick (1989) for example, mentions *construct-irrelevant variance* (when one or more irrelevant constructs is being assessed in addition to the intended construct) and *construct underrepresentation* (when not all critical components of the intended construct are included in the assessment) as the two major threats of the validity of PA. According to Lane and Stone (2006) in particular the variety of sources of construct-irrelevant variance the raters bring to the rating process are cause of concern (e.g. rater attention to irrelevant features, rater tendencies to severe or lenient rating, or divergent interpretation of the content standards). They say a pilot study provides an opportunity to evaluate the quality and comprehensiveness of the content and processes being assessed and to trace potential issues of bias in task language and context. Tasks and scoring procedures may be reviewed and modified a number of times prior to and after being piloted, improving the quality of the final product. Therefore, in April 2011 a pilot of the UHTRUST took place in both Utrecht and Hamburg. In this study the validity of this pilot assessment is examined.

1.1.4 Aim and research questions

This study is part of a larger research project on the development of the UHTRUST. The aim of this study was to deliver the initial validity evidence of the pilot version of the PA. This validity evidence consists of an *interpretative argument*, containing the proposed assessment *interpretation and use* of the test scores, and a corresponding *validity argument*, containing different kinds of *backing* for the proposed interpretation and use. This backing was gathered during the development, implementation and evaluation of the UHTRUST pilot version. The used validation method is in accordance with the *argument-based approach to validity* as provided by Kane (2006), which is considered the standard approach for test validation.

The main goal here was to acquire an explicit, coherent, and plausible argument for validity. This can be done by laying out the network of *inferences and assumptions* leading from the test performance to the conclusions to be drawn. This study had a confirmative approach; it sought justification for the validity of the UHTRUST pilot version. Kane (2006) emphasizes that this advocacy role is legitimate in the early stages of test development. Objective appraisal is only appropriate and possible in a more mature phase, in which the test is challenged and expected to stand up to criticism. However, this study also intended to critically evaluate the validation evidence as far as possible during the development stage of a test, and make recommendations for validity improvement after the pilot. The research questions is:

(1) What is the validity of the UHTRUST pilot version?

The theoretical framework below provides insights into the broad context of this study. General information about PAs and validity, and the argument-based approach to validity is given.

1.2 Theoretical framework

1.2.1 Performance assessment and validity

1.2.1.1 Definitions

Kane, Crooks and Cohen (1999) stated that “the defining characteristic of a performance assessment is the close similarity between the type of performance that is actually observed and the type of performance that is of interest” (p. 7). They meant that the high-fidelity tasks in a PA can easily be translated to expected performance in the real-world. A PA makes it possible for an assessor to truly see the student perform in an authentic environment. The inference can be made from a high score for a task in a PA, to a high level of proficiency in a similar task in practice. Lane and Stone (2006) describe PA as an assessment that requires students to perform an activity or construct an original response. During the assessment students are stimulated to apply their problem solving skills in relatively novel real-world situations. Multiple solutions or strategies are possible and the duration of the assessment can range from several minutes to several days or more.

Gipps (1994) stated that while not all PAs are authentic, it is difficult to imagine an authentic assessment that is not a PA. According to her an authentic assessment is a special case of a PA. Meyer (1992) suggested that test developers should specify in which respects the assessment is authentic; the stimulus, task complexity, locus of control, motivation, resources, conditions, criteria, or the standards? The list is almost infinite but the point is that test developers should clarify their definition and address the question ‘authentic to what?’

1.2.1.2 The development of a PA

Lane and Stone (2006) describe the design of a PA as an iterative process. They stated that the generalizability across variations in tasks, settings, and examinee groups can be reinforced by using a construct-centred approach. Firstly, the construct is defined and the *most critical*

features of the target domain that need to be assessed are identified. Based on this construct description the performances or behaviors that should be elicited by the assessment can be identified. Next, this construct is used as a guide in the actual task development as well as the specification and development of the scoring criteria and rubrics. In this way, generalizability is enhanced, since the test developer is forced to pay attention to the possibility of construct-irrelevant variance and construct underrepresentation. Alignment among the defined construct, task, and scoring methods is ensured.

1.2.1.3 *Validity and reliability issues in PAs*

PAs bring along specific problems concerning validity and reliability. An important difference with traditional assessments is that in PAs *assessors* are used who judge the performances shown by the respondents during the assessment (Bakker, 2008). As a result of the inclusion of human judgment specific threats arise. As mentioned earlier, Lane and Stone (2006) say that a major concern with rater-scored assessment is that raters bring a variety of potential sources of construct irrelevant variance to the rating process, such as rater attention to irrelevant features, or rater tendencies to severe or lenient rating. According to Nijveldt (2007) assessors often show a tendency to consider only confirmatory evidence for their initial interpretations, as opposed to also considering counterevidence or alternative interpretations. Assessors bring unique characteristics, experiences and schemata to the rating process, and it appears to be difficult for assessors to exclude *biases* stemming from their personal backgrounds during the judgment process, and to prevent selective observations (Bakker, 2008; Govaerts et al., 2007; Sterkenburg, Barach, Kalkman, Gielen & Ten Cate, 2010).

Furthermore, in PAs *open-ended* and *complex* tasks are used, and respondents can react to those open-ended tasks in very different ways. It is hard for assessors to score these performances in a consistent way (Gipps, 1994; Moss, 1994). This is a problem, since a certain degree of reliability is important for large-scale assessments that have high stakes associated with their use (Kane, 2006; Knight, 2002). The scoring process of such tasks asks for much personal interpretation, and some degree of subjectivity may enter into the scoring (Bakker, 2008; Govaerts et al., 2007; Lane & Stone, 2006; Van der Schaaf & Stokking, 2008).

Another threat concerning the validity and reliability of PAs is *the nature of the assessment tasks* (Bakker, 2008). In PAs a high score on one task does not necessarily mean a high score on a different task, even when the tasks are from the same domain (Lane & Stone, 2006). According to Gipps (1994) performances on open-ended and complex tasks are highly *task-specific*; performances on tasks that appear to be similar, will in fact only be moderately similar. Little is known about the actual causes of the divergent performances of respondents on different assessment tasks (Bakker, 2008).

The selection of *representative samples of assessment tasks* is another important issue in PA (Bakker, 2008). The authentic character of PA tasks tend to make these tasks very time consuming, and therefore only a relatively small number of tasks can be included in the PA (Kane, 2006; Lane & Stone, 2006). Because of this small number of performance tasks, unique features of a task have a much bigger effect on the total score than for example a multiple-choice test with a large number of items, where students' individual reactions to specific tasks tend to average out (Kane, 2004). A small number of tasks in a PA makes it

difficult to establish a sample of assessment tasks that is representative to the universe of generalization (Kane, 2006).

Finally, *scoring criteria* can become a threat to the validity and reliability of a PA when they are not formulated precisely (Bakker, 2008; Solomon, Szauter, Rosebraugh & Callaway, 2000). If criteria are formulated too narrowly, there is a risk that assessors get lost in the atomistic details and miss the essence of the performance. If criteria are formulated too broadly, then it is difficult for assessors to apply these criteria consistently.

1.2.1.4. Some solutions to validity and reliability issues in PAs

The attainment of acceptable levels of reliability in PA and other new forms of assessment is a notorious problem (Van der Schaaf, Stokking & Verloop, 2005). Attempts for improvement have focused primarily on standardization and objectivity of measurement by adjusting assessment instruments, rating scale formats and enhancing raters' accuracy and consistency through rater training. Govaerts and colleagues (2007) argue that this psychometric approach tends to ignore the role of the assessment context and the unique individual experiences and schemata raters bring to the rating process. Van der Schaaf and colleagues (2005) therefore stated that the percentage of exact agreement in scoring between raters can only be used as an *indicator* of reliability. Systematic differences in rater background and knowledge structures might make it more realistic to alter the reliability demands. Perhaps, it is fair enough when each rater is consistent in his or her rating according to his or her own interpretations of the standards. If so, validity purposes will make it more and more important to capture cognitive processes underlying raters' judgment.

This does not mean that standardization measures in order to enhance objectivity and reliability in PAs can be omitted. Hawkins and colleagues (2009) stated that an assessment score that tells the user more about the evaluator or the setting than it does about the examinee, does not become useful because the context in which it was produced was authentic. They argue that *rater training* should be used in order to ensure that raters are aware of typical rater-errors and performance standards are consistently applied. Nijveldt (2007) also stated that it is important for assessors to be aware of their own judgment process and the specific threats to validity in these processes. She distinguished several indicators that secure a valid judgment, i.e. all relevant evidence is considered; both confirmative evidence and counterevidence are considered; there is no idiosyncratic weighing of evidence; and all conclusions are supported by clear argumentation which also draws upon the original data. No extraneous, irrelevant criteria should be added. Different training formats for rater training exist. More interactive training formats, involving practice ratings and feedback, are often found most efficient in reducing leniency error and improving rating accuracy (Hawkins et al., 2009).

Kane (2006) further stated that the use of a large number of assessors in PAs contributes to more reliable scoring, and that the use of multiple assessors per examinee decreases the influence of personal biases of the individual assessors. He also argues that the reliability of the assessment outcomes can be enhanced by standardization of assessment tasks. Standardization reduces the openness of the tasks, enhancing more homogeneous reactions from the respondents to the tasks, and reducing problems with task specificity. Finally, the representativeness of the sample of assessment tasks can be increased by raising

the number of included assessment tasks. However, in practice this measure is only attainable when time and money allow it.

1.2.2 The argument-based approach to validity

1.2.2.1 Definitions

According to the *Standards for Educational and Psychological Measurement* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) “validity refers to the degree of which evidence and theory support the interpretations of test scores entailed by purposed used of tests” (p. 9). Earlier, Messick (1989) stated that validity involves the meaningfulness, appropriateness, and usefulness of test score inferences. Even though validity theory is quite sophisticated, the theory is also quite abstract, and does not provide clear guidance for validation (Kane, 2004, 2006). Often, this lack of guidelines resulted in the tendency of test developers to be somewhat opportunistic in the choice of validity evidence, but it also resulted in frustration of willing test developers, who did not know where to start and when enough evidence was enough (Kane, 2004; Chapelle, Enright, & Jamieson, 2010). By proposing the argument-based approach to validity Kane (2004) “intended to improve this situation by providing a methodology or technology for validation” (p. 136).

1.2.2.2 Two stages within the validation process

The argument-based approach involves two stages (Kane, 2006). The first stage is the *development stage*, in which the assessment procedures are developed, and the proposed interpretation and use are specified by means of an interpretative argument. As far as possible during this stage, the corresponding validity argument provides accurate backing for the interpretative argument. The purpose of this stage is to make implicit inferences and assumptions explicit, so it becomes “more difficult to fool ourselves into believing we have validated an interpretation and use” (Kane, 2004, p. 143). The product of the development stage is an explicit, coherent interpretative argument linking test performances to the proposed interpretations and uses.

Kane (2006) argues true critical evaluation during the development stage is premature. Therefore, the proposed interpretative argument is critically evaluated in the second stage, i.e. the *appraisal stage*, when the measurement product is finished. The evaluation begins with a review of the interpretative argument as a whole to determine if it makes sense. Then, all individual assumptions are checked; the easy ones and the most questionable. Inspired by Toulmin (1958), Kane (2004) encourages the comparison of the argument-based approach with a dialog or debate between the proposer of a claim and the challenger. Evidence for and against are weighted carefully. Hidden assumptions and possible alternatives have to be identified. If the test is not able to withstand the challenges, the assessment procedure or the argument has to be revised or abandoned. If the test can withstand the challenges, confidence in the claims increases. However, even when at the end of this iterative process the interpretative argument is found to be coherent and all its inferences and assumptions turn out to be plausible, conclusions may be overturned in *exceptional* cases, if they are contradicted by reliable evidence (e.g. plagiarism, a noisy test

environment, or rater bias). Kane (2004, 2004) calls this the *ceteris paribus* assumption, i.e. the “all else being equal” assumption.

1.2.2.3 *The structure of an argument for validity*

Since the UHTRUST project is in its the development stage, it is appropriate to focus on the construction of an initial argument for validity. An argument of validity usually exists of four major inferences: (1) *scoring*; (2) *generalization*; (3) *extrapolation*; and (4) *interpretation* (Hawkins, Katsufakis, Holtman & Clauser, 2009; Kane, 2004; Kane, 2006). Each inference involves an extension of the interpretation; the conclusions or claims of earlier inferences serve as starting points for later inferences.

The *scoring* part of the argument requires evidence regarding the appropriateness of the assessment conditions, data collection, scoring procedure and the justification for the scoring rules. The initial datum is likely to be the record of a performance. Backing for this part of the argument generally consists of testimonies of the participants, interrater reliability data, and procedural evidence (e.g. a pilot or expert judgment).

The *generalization* part of the argument requires evidence that the observed score, coming from the task sample used in a particular test, is generalizable to a broader domain, i.e. the *test domain*. The test domain is the *universe* of all possible tasks or items the test developer can use in his design. The examinee’s test score does not change, but the interpretation of the score is broadened from a specific set of performances to a claim about the candidate’s *expected* performance in the test domain. To do so, it is important that an examinee’s score is stable and does not change dramatically if he or she was retested on a different *occasion*, scored by a different *rater*, and examined with a different sample of test *tasks* or items. Kane (2006) emphasises that the sampling error has to be controlled by using a large enough sample from the test domain and by making sure that the test tasks or items are representative of the *universe of generalization*. The backing usually includes sampling theory and/or expert judgment (to answer the question whether or not the sample was large enough and representative of the universe), and reliability or generalizability studies (which provide estimates of standard errors and confidence intervals of measurement and therefore puts limits on the precision of estimates of the universe score, i.e. *qualifiers*).

In the *extrapolation* stage of the argument once again the interpretation is extended to a broader domain, i.e. from the *universe of generalization* to the *practice domain* or *target domain*. This requires evidence of the link between the data collected in the test and the behaviours of interest in the real-world (Hawkins et al., 2009). This can be done by demonstrating the overlap between the skills measured by the test and those needed in practice. Extrapolation has two major threats: it can be undermined by construct irrelevant variance, and construct underrepresentation. Backing in the extrapolation stage can consist of expert judgment, analytic evidence (aimed at making the test as representative of the target domain as possible), and empirical evidence (examination of the relationship between observed scores and other scores or variables associated with the target domain).

The fourth stage of the argument is the *interpretation* or *decision* stage. Here, a final and appropriate conclusion is drawn: what implications logically result from the observed skill level of the examinee? Backing aims at the justification of the *decision rule* and often consists of an analysis of positive and negative *consequences* resulting from the decision,

relative to those for alternative decision procedures. To do so, the appropriateness of the *performance standards* and *cut-off score* are investigated and attention is paid to possible *irrelevant method variance* (Hawkins et al., 2009; Kane, 2006).

2. Method

2.1 Participants

2.1.1 Test developers

2.1.1.1 Educational researchers

During the development and evaluation process of the UHTRUST pilot version, nine Dutch and two German educational researchers contributed to different aspects of the design. Among others, the delineation of the competences of interest, and the development of the assessment content, were important aspects of the test that were designed and arranged by these researchers. The contributing researchers all influenced the overall test validity and consequently are participants in this validation study.

2.1.1.2 Experts

Also, 35 experts were consulted during the development and evaluation process of the UHTRUST pilot version. Eight Dutch physicians contributed to the task development process, an American statistical expert was consulted during the development and evaluation of the scoring criteria, scoring rubric and scoring process, and 18 Dutch and eight German physicians participated in a Delphi study, in which the critical features that needed to be assessed were identified.

2.1.2 Test participants

The validity of the pilot assessment was also determined by the persons who participated in the assessment, i.e. the test respondents, assessors, and simulation patients (SPs). For example, the appropriateness of their personal backgrounds may have a positive or negative influence on the validity of the UHTRUST pilot version.

2.1.2.1 Test respondents

Although the ultimate target population of the UHTRUST project is binary (VI-graduates and non-VI graduates), a more heterogeneous group of respondents was used during the pilot assessment. The respondents of the pilot assessment consisted of 4 groups: junior and final year medical students of the UMCU (Utrecht) and the UKE (Hamburg), regional residents, and experienced general practitioners (GPs) were approached with information on the study and a request for participation. Because of low response rates the personal network of the main researcher of the UHTRUST project was also used to engage people.

In *Utrecht* a total of 12 persons were willing to participate in the pilot assessment: three junior medical students, four final year students, three GP residents, and two experienced GPs. Ten participants were female, and two participants were male.

In *Hamburg* a total of eight persons were willing to participate in the pilot assessment: two junior medical students, three final year students, and three residents (one nephrology, one general internal medicine, and one experimental oncology resident). All participants happened to be female.

All respondents of the pilot assessment participated voluntarily and signed an informed consent form. In return for their participation, respondents received personal feedback about their performances on the test, and gift coupons.

2.1.2.2 Assessors

Each respondent of the UHTRUST pilot version in *Utrecht* was evaluated by at least four, and (when possible) sometimes five trained assessors, namely two experienced nurses and two or three experienced physicians. In *Utrecht* nurses and physicians of the UMCU were approached with information on the study and a request for participation. Five nurses and ten physicians were willing to participate. The Dutch nurses all had an average working experience of 14 years. Their average age was 35 years, and all nurses happened to be female. The Dutch physicians all were specialists in various medical disciplines (internal medicine, anesthesiology, rheumatology, pediatrics, urology, family medicine and thoracic surgery). They had an average working experience of 23 years and an average teaching experience of nine years. Their average age was 52 years, and three physicians were female and six physicians were male. On the assessment day one Dutch physician canceled his participation because of illness, so one respondent only was evaluated by three assessors: two nurses and one physician.

In *Hamburg* each respondent was evaluated by at least three, and (when possible) sometimes four, trained assessors, namely one experienced nurse and two or three experienced physicians. Nurses and physicians of the UKE were approached with information on the study and a request for participation. Here, three nurses and nine physicians were willing to participate. The German nurses had an average working experience of 24 years. Their average age was 40 years, and again all nurses happened to be female. The German physicians all were specialists in various medical disciplines (surgery, internal medicine, neurology, cardiology, gynecology, and pediatrics). They had an average working experience of 14 years and an average teaching experience of five years. Their average age was 41 years, and three physicians were female and six physicians were male.

On the day prior to their participation, all of the aforementioned assessors followed a frame-of-reference training course. During this half-day training course the procedure to assess the respondents ability to take responsibility for unfamiliar clinical situations was introduced and the scoring criteria and corresponding performance levels were plenary discussed. The nurses also received information about the disruptions they had to act out during the pilot assessment.

All assessors of the pilot assessment participated voluntarily. In return for their participation the assessors received gift coupons. The physicians also received accreditation points for their attendance to the training course.

2.1.2.3 *Simulation patients*

Although in the ultimate assessment medical graduates will encounter 5 SPs, respondents of the pilot assessment had 6 clinical encounters, i.e. with 1) a mother of a 5 year old girl with fatigue and stomach ache; 2) a 53 year old man with progressive fatigue and haemoptysis; 3) a 58 year old woman with stomach ache; 4) a 62 year old woman with a stomach ache; 5) a 65 year old woman with speech-, chewing-, and swallowing problems and her spouse; and 6) a 33 year old man with hypertension, anaemia, and epilepsy.

In *Utrecht* seven professional actors were hired and trained to portray these six patient scenarios in a standardized way (the enactment of scenario five required the presence of two actors). In *Hamburg* 14 professional actors were hired and trained (the double amount), so the actors could take turns. During their training the actors learned to portray the scenarios in a realistic and standardized way. All actors had experience as SP.

2.2 Instrumentation

2.2.1 *Instrument used during the development of the argument for validity*

Preceding the development of the pilot assessments' argument for validity, various theoretical and empirical studies related to the argument-based approach to validity were thoroughly examined, i.e. Bakker (2008); Chapelle and colleagues (2010); Hawkins and colleagues (2009); Hawkins and Holmboe (2008); Kane (2004); and Kane, (2006). Based on these studies a theoretical framework for an argument for validity was developed (see Table 1). The framework illustrates the four major inferences that are associated with the argument-based approach to validity and their underlying assumptions. The framework served as a blueprint during the writing process of the argument. This was done to prevent that the argument for validity would become incoherent, opportunistic, and/or never-ending.

2.2.2 *Instruments used to collect validity evidence during the pilot assessment*

During the pilot assessment, three instruments were used to collect validity evidence. These three instruments were: an analytic scoring rubric, a holistic scoring rubric, and evaluation forms.

2.2.2.1 *Analytic scoring rubric*

During the assessment day, all assessors gradually filled in a ten-page *analytic* scoring rubric. Each page on the scoring rubric contained the rubric of one of the ten critical features as agreed upon by the experts in the aforementioned Delphi study. These features were used as scoring criteria. The assessors were asked to fill in the analytic scoring rubric according to a detailed *scoring procedure*. According to this scoring procedure, it was important that the assessors first observe the respondents and collect evidence pertaining to their performance levels. The assessors used the open spaces on the analytic scoring rubric to write down all this evidence (i.e. relevant behaviors). Eventually, when all relevant evidence was collected the assessors used these separate pieces of evidence to form an overall judgment about the respondents' performance on a scoring criterion. The analytic judgments were given on a 5-point scale, ranging from 'Poor' (score 1) to 'Excellent' (score 5). A high analytic score

Table 1. Framework for an Argument for Validity: The Four Major Inferences of Kane's Argument-Based Approach to Validity and their Associated Assumptions

<p><i>Inference 1, scoring:</i> from the observed performance to the observed score</p> <ul style="list-style-type: none"> - <i>Assumption 1.1: The assessment conditions are appropriate;</i> - <i>Assumption 1.2: The scores are recorded accurately;</i> - <i>Assumption 1.3: The scoring criteria are appropriate and acceptable;</i> - <i>Assumption 1.4: Reliable and valid scoring of the performance by the assessors.</i>
<p><i>Inference 2, generalization:</i> from the observed score to the expected universe score</p> <ul style="list-style-type: none"> - <i>Assumption 2.1: The scores are stable and random error due to different occasions, raters and tasks is controlled;</i> - <i>Assumption 2.2: The sample of observations is representative of the universe of generalization.</i>
<p><i>Inference 3, extrapolation:</i> from the universe score to the expected level of skill in the target domain</p> <ul style="list-style-type: none"> - <i>Assumption 3.1: The universe score is related to the level of skill of the graduate in the target domain;</i> - <i>Assumption 3.2: There are no systematic errors that are likely to undermine the extrapolation.</i>
<p><i>Inference 4, interpretation:</i> from the level of skill in the target domain to the test interpretation</p> <ul style="list-style-type: none"> - <i>Assumption 4.1: All assumptions are defensible with accurate and plausible evidence;</i> - <i>Assumption 4.2: The data acquired by the assessment can be used for the intended purposes.</i>

indicated that the respondents behavior during the day, concerning one specific scoring criterion (e.g. criterion 4: teamwork and collegiality), had led to positive consequences for his SPs (or other fictional patients). On the rubric assessors also indicated on a 3-point scale how certain they were of their judgment (1 = 'Uncertain'; 2 = 'Partially certain'; and 3 = 'Certain').

2.2.2.2 Holistic scoring rubric

At the end of the assessment day, the assessing *physicians* also filled in a *holistic* scoring rubric. Considering the *total* amount of collected evidence on the analytic rubric, physicians were asked if they would entrust the respondents to take responsibility for ten other unfamiliar clinical situations. For example, physicians were asked if they would entrust the respondent to 1) render emergency assistance to a patient with acute decompensated cardiac failure; 2) handle a request for organ donation; or 3) bring bad news to a patient. These holistic scorings enabled a more direct measurement of the respondents' ability to take responsibility for

unfamiliar clinical situations than the analytic scorings. Each of the ten situations required the use of *several* of the aforementioned *analytic* features. The ten new unfamiliar situations were specified on the holistic scoring rubric by means of short written cases. The physicians made their holistic entrustment decisions on 7-point scales, ranging from ‘I would not want this trainee to take this responsibility in the first place’ (score 1) to ‘I would trust this trainee to supervise another trainee carrying out this task’ (score 7).

Both the analytic and the holistic scoring rubrics provided some of the backing for the *scoring* inference (assumption 1.3 and 1.4 in Table 1).

2.2.2.3 Evaluation forms

At the end of the assessment day all 67 test participants (respondents, assessors, SPs) and educational researchers (and other staff members who helped coordinating the assessment day) filled out a form on which the pilot assessment was evaluated. The evaluation forms aimed to gather evidence concerning procedural validity of the UHTRUST pilot. The participants were asked to indicate to what extent they agreed with various statements. This was done on a 5-point scale, ranging from ‘I completely disagree’ (score 1) to ‘I completely agree’ (score 5). ‘*The assessment day was well organized*’ is an example of a statements that was presented to all test participants and staff members. Depending upon their role, test participants were also asked to give their opinion on more tailored items. For example, assessors were asked if they thought that they had received sufficient time to fill in the scoring rubrics. A total of 11 items were evaluated. One of these items was an open ended question.

The evaluation forms provided some of the backing for the *scoring* inference (assumption 1.1, and 1.4 in Table 1) and the *extrapolation* inference (assumption 3.1, and 3.3 in Table 1).

2.3 Procedure

The writing of the argument for validity was an iterative process. All pieces of validity evidence had to be collected and arranged in a way that did justice to Kane’s argument-based approach to validity. The theoretical framework of an argument for validity (Table 1) helped to do so.

In order to collect the validity evidence, close contact was held with all test developers. During the development of the pilot assessment they were asked to clarify their considerations and to check the veracity of the parts in the argument that concerned their decisions or designs. Often, these decisions and designs were altered, which meant that (parts of) the argument had to be rewritten as well. Most of the procedural validity evidence was developed and gathered during this stage.

During the implementation of the pilot assessment empirical evidence and data with respect to the internal validity of the design was collected by means of the scoring rubrics and evaluation forms. This data greatly influenced the validity evidence that was developed and gathered during the development stage of the pilot assessment, because it gave insight into the actual quality of the design. The internal validity evidence determined the strength of the claims made in the argument for validity.

2.4 Analysis

During the data analysis all available data was used to design a coherent and plausible argument. To do so, the data was linked to one or more of the four major inferences in the argument for validity (see Table 1).

2.4.1 Analysis of the data gathered during the development of the pilot assessment

As mentioned above, most of the procedural evidence was developed and gathered during the development stage of the pilot assessment. During this stage test developers based their most important decisions on theory and expert judgment (i.e. the expert judgment of the aforementioned physicians and statistical expert). With the help of the test developers, the most important theory was processed as backing in the pilot assessments' argument for validity. Because most experts were consulted in a relatively informal setting, test developers were asked to recall the exact numbers and backgrounds of all contributing experts. This information was also processed as backing in the argument for validity, because such information increased the credibility of the argument. Exceptions to this informal setting were the experts that were consulted by means of a Delphi study.

2.4.2 Analysis of the data gathered during the implementation of the pilot assessment

2.4.2.1 Scoring rubrics

During the pilot assessment the assessors produced two types of scores: analytic and holistic scores. The analytic scores were used to calculate a third type of scores: the total analytic mean scores. The holistic scores were used to calculate a fourth type of scores: the total holistic mean scores.

The total analytic mean scores and the total holistic mean scores were used to find part of the evidence for construct validity: by means of ANOVAs the effect of medical experience on test scores were calculated. If the pilot assessment is valid, respondents with the most experience with coping with unfamiliar clinical situations (i.e. the residents and the experienced GPs) should perform significantly better on the test than respondents with the least experience (i.e. junior and last year medical students).

Evidence for construct validity was also gathered by means of principal component analysis. It was verified whether the analytic scores on the ten scoring criteria, and the holistic scores on the ten new unfamiliar scenarios were driven by the supposed underlying construct: the ability to take responsibility for unfamiliar clinical situations. If the pilot assessment is valid, only one component should underlie each type of score.

Next, the reliability of the analytic and holistic scores were analysed by means of the Cronbach's alpha test.

Because the analytic criteria and holistic scenarios were designed to measure the same construct, Pearson correlation coefficients were calculated in order to check whether relationships between holistic and analytic scores existed.

All assessors indicated how certain they were about their analytic judgments. These data were used to check whether some scoring criteria were found harder to judge than others.

The agreement between rater pairs on the analytic and holistic scores was calculated by means of Cronbach's alphas (jury alphas). As an additional indication of the consistency of raters' judgments, it was verified whether there were statistically significant differences between the total analytic mean scores and the total holistic mean scores.

2.4.2.2 Evaluation forms

Data from the evaluation forms were analyzed using descriptive statistics: mean scores (M), standard deviations (SDs), frequencies and percentages of the items and groups of interest were calculated.

3. Results

3.1 Inference 1: scoring

“It is reasonable to assume that administration conditions have a non-negligible effect on respondent performance” (Cohen & Wollack, 2006, p. 357).

It is important that a high or low score on the pilot assessment corresponds with the respondents' behavior during the pilot assessment and the respondents' true capacity on the assessed trait. In the first part of this argument the score-equating strategies designed to enhance stability and equal opportunities for all respondents to show their abilities - regardless of the administration occasion, assigned assessors, and simulation patients - are described and evaluated. In Table 2 a summary of the scoring inference is presented.

3.1.1 Assumption 1.1: The assessment conditions are appropriate

3.1.1.1 Standardization of the assessment conditions

The UHTRUST pilot version was administered on two different occasions, on two different locations, and with two different groups of respondents, assessors, and simulation patients (SPs); once Dutch respondents were judged by Dutch physicians and nurses at the UMC Utrecht, and once German respondents were judged by German physicians and nurses at the UKE Hamburg. Language barriers and physical distance made it impossible to test all respondents on the same occasion, on the same location, and with the same group of assessors and SPs. Joint examination in English was considered but rejected because of the extra systematic error this would bring about; it would have had a negative effect on the level of authenticity of the performance assessment (PA). In order to promote fairness between the two administration occasions and control random error the observations of the PA were made under standardized assessment conditions (see Kane, 2006; Cohen & Wollack, 2006). Preceding the pilot assessment the standardized directions, conditions of administration, and scoring were clearly prescribed by the developers of the test. Directions included things as time limits, how to construct a valid response or judgment, what ancillary materials were allowed, and how much help the assessors and SPs could be expected to provide. During the assessment all respondents were assessed with the same clinical content, and under the same

Table 2. Summary of the Scoring Inference

Assumptions underlying the scoring inference	Warrants licensing the assumptions
1.1: The assessment conditions were appropriate	<p>The assessment <i>conditions</i> were <i>standardized</i>, so respondents were provided with equal opportunities to show their abilities, and test scores could be compared to one another.</p> <p>A detailed <i>planning</i> was defined to ensure a smooth running of the assessment day. This planning was evaluated.</p>
1.2: The scores were recorded accurately	<p>All respondents were judged by <i>multiple trained assessors</i>. Depending on their role, the assessors judged the respondents behavior on a holistic and/or analytic scoring rubric.</p> <p>Assessors were urged to follow a systematic and transparent <i>scoring procedure</i>. This reduces the risk of invalid and unreliable judgments.</p> <p><i>Security measures</i> were taken to prevent loss of data and to protect the tests' integrity.</p>
1.3: The scoring criteria were appropriate and acceptable	<p>A panel of informed experts <i>agreed upon</i> the content and language of ten <i>scoring criteria</i>. These criteria were used to develop an <i>analytic scoring rubric</i>.</p> <p><i>Global rating scales</i> were used on the analytic scoring rubric. They were expected to be more feasible, equally reliable, and more valid than detailed (dichotomous) behavioral checklists.</p> <p>Experts developed ten <i>EPA scenarios</i> that were used to develop a <i>holistic scoring rubric</i>.</p> <p><i>Construct validity</i> was statistically evaluated.</p>
1.4: Reliable and valid scoring of the performance by the assessors	<p>The <i>sample</i> and <i>selection procedure</i> of the assessors was acceptable.</p> <p>During a <i>frame-of-reference training</i> assessors attempted to reach shared understanding of the content and performance standards. Assessors were also informed on how to avoid typical rater-errors.</p> <p>The <i>quality</i> of the scoring criteria, EPAs, and scoring procedure was checked; on the analytic scoring rubrics a lot of scores were missing, in particular criteria 5 and 7 were often neglected.</p> <p>The <i>consistency</i> between most assessor groups and pairs on the holistic scoring rubric was acceptable.</p>

conditions. The pilot assessment consisted of three phases. In the *first phase*, respondents had short clinical encounter with six trained SPs. In the *second phase* respondents were asked to make a management proposal for each of the six unfamiliar cases. During their period of study and thought, all respondents faced standardized disruptions, i.e. phone calls and face-to-face questions from ‘colleagues’ on various work related subjects. In the *third* and final *phase*, the respondents reported on their management proposals. These reports were presented orally and in written form; the mode of presentation was the same for all respondents. For the three phases respondents received respectively, one hour and 25 minutes, four hours, and 40 minutes to complete their tasks.

The human element and the open-ended nature of the assessment tasks of the UHTRUST pilot administration only made it possible to work in partially standardized assessment conditions. In this kind of assessment it is difficult to discern all the potential threats to standardization, “including those associated with SP portrayal, unanticipated student reactions to the scripted SP responses, and case irregularities” (Hawkins & Holmboe, 2008, p. 105). The pilot assessment provided an opportunity to act out and critically evaluate all score-equating strategies.

3.1.1.2 Assessment planning

To ensure a smooth running of the pilot assessment a detailed planning was defined. This planning was not purely based on commonsense, but was complemented by empirical and theoretical studies containing useful advice about successfully running PAs similar to the UHTRUST pilot version (int.al., Boursicot & Roberts, 2005; Cohen & Wollack, 2006; and Hawkins & Holmboe, 2008). Besides detailed time schedules for the respondents, assessors, SPs, and researchers, the planning included descriptions of the necessary practical and logistical arrangements.

On the evaluation form both Dutch and German participants (N = 84) were asked if the day was well organized. The mean score on this item was 4.11 (SD = 0.44). On a 5-point scale 85% of the participants scored this item ≥ 4 . A notable fact was that one group seemed far less satisfied with the organization than other groups: the mean score of the German staff members (N = 7) on the item was 2.30 (SD = 0.41). They indicated that they had not received clear instructions beforehand. This obscurity resulted in a violation of the specified procedures: German respondents received several minutes more for the encounters with the SPs in the first phase of the pilot assessment than allowed. This did not enhance the fairness of the pilot assessment and detracted the standardization measures.

3.1.2 Assumption 1.2: The scores are recorded accurately

3.1.2.1 The recording of the test scores by multiple assessors

Each respondent of the UHTRUST pilot version was evaluated by multiple trained assessors, i.e. nurses and physicians (see method). The assessors acted out different roles and observed different parts of the respondents’ activities during the day. The nurses were present during the second phase of the assessment. Here, they were allowed to interact with the respondent about the cases, ask (scripted) questions and observe all activities. During the day each

respondent also was linked to a physician, who took on the role of the respondents' *absent* supervisor. This physician was asked to perform this role like he or she would have done in real life. As in real life, respondents were allowed to talk with their supervisor about the cases. Because of the supervisors' absence, these consultations were mostly done by phone. During the day the supervisor was shadowed by a second physician, who monitored all interactions between the respondent and the supervisor. This physician was only allowed to talk with the respondent during the third phase of the pilot assessment. Interaction between the two paired physicians was not allowed, because this would influence their personal judgments. On some occasions a third physician was attending the respondents' final report on his day on the ward. On one occasion a supervisor was not shadowed by a second assessor (see method).

During the assessment day physicians were responsible for the scoring of three respondents, and nurses for the scoring of four respondents. Based on the impressions and observations of the process during the day and/or the final product at the end of the day, the nurses and physicians evaluated the respondents' clinical skills on the holistic and/or analytic scoring rubric. These scoring rubrics were designed to ensure that all critical features of coping with unfamiliar clinical situations were systematically considered by all assessors. The nurses and physicians were expected to score their observations of the respondents' performances according to a detailed *scoring procedure* (see method). The use of such a systematic and transparent scoring procedure reduces the risk of invalid and unreliable judgments (Bakker, 2008; Kane, 2006).

3.1.2.2 Accurate data collection

To prevent loss of data or mixing up of data various security measures were taken. At the beginning of the day all respondents, assessors, and SPs checked-in. At the end of the day all participants checked-out and researchers made certain that all documents that should have been received were received. The researchers who collected the documents also checked whether all required information for accurate data processing, scoring, and reporting was present and readable. After entering the data on the computer the integrity of the imported data was checked before scoring was done.

3.1.2.3 Test security

The UHTRUST pilot administration was intended for research purposes only. Since there were no official positive reinforcements or harmful consequences for those who had a high or low score on the test, and because participation was voluntarily, it was assumed that the respondents' motivation to cheat was low. Individual or collaborative cheating by assessors, who can artificially increase respondents test scores on their scoring rubrics, could be motivated by pride and competitive feelings between German and Dutch assessors. The effects of this kind of cheating were expected to average out. However, to protect the assessments' integrity, prior to the pilot assessment all assessors received training in which the impact of rater violations was explained.

When respondents have access to the test tasks prior to the test administration, the scores of these respondents do not accurately reflect their ability levels. This can happen when respondents share test content with respondents who have yet to be tested, when information leaks out or is stolen early in the testing window, or when respondents are already exposed to

the test task in previous assessments (Cohen & Wollack, 2006). Language barriers, physical distance, and lack of motive made it unlikely that German respondents, who were tested several days before the Dutch respondents, shared information with examinees in Holland. The patient scenarios were developed de novo in the months prior to the pilot assessment by various educational researchers and medical experts (see method and inference 2). These scenarios were based on real patients, but were tailored to fit the proposed assessment use and target population of the UHTRUST project. The scenarios were put in trial for the first time during the pilot assessment. Premature exposure to the assessment tasks was therefore unlikely.

3.1.3 Assumption 1.3: The scoring criteria are appropriate and acceptable

3.1.3.1 The analytic scoring criteria

The ability to cope with unfamiliar clinical situations on the ward is a diverse and complex construct that is not easy to capture. In order to measure this broad construct appropriate scoring criteria had to be identified for the UHTRUST. Scoring criteria (or content standards) indicate what respondents should know and be able to do, and they guide the assessors' judgment about the quality of a respondents' performance (Gipps, 1994).

For the UHTRUST ten *critical features* were identified for assessing the respondents' overall clinical skills while dealing with unfamiliar situations. These features were used as scoring criteria on an analytic scoring rubric. A high ability to cope with unfamiliar situations was revealed by a high score on these ten scoring criteria; the respondents' overall clinical skills were acceptable (or even excellent), despite of the unfamiliar clinical situations the respondent was confronted with.

3.1.3.1.1 The development of the scoring criteria

The development of the analytic scoring criteria was split into two stages. In the first stage, published criteria and standards of three acknowledged (inter)national physician competency frameworks and three related empirical studies were screened by an educational researcher to establish a preliminary list of 25 potential criteria. A criterion was only selected as potential criterion when the formulation of the criterion was accurate and suited the final target population. In the second stage, a Delphi method was used to produce a tailored set of criteria. Both Dutch (N= 18) and German (N= 8) physicians were asked to participate as panellists in the Delphi-study. These physicians were selected based on their extensive experience in postgraduate training, which made it possible for them to make accurate judgments about the level of proficiency that could be expected from the projects' target population. In three rounds the experts scrutinized the preliminary list and judged and revised the criteria. In the first round 14 Dutch experts judged the criteria on a 7-point scale ranging from low relevance (score 1) to high relevance (score 7). They also commented on the quality and comprehensiveness of the individual criteria. In the second round this process was repeated with 15 Dutch experts and consensus about content and language was reached. The renewed list of 25 criteria was translated in German. In the final round 8 German and 8 Dutch experts ranked the criteria. This method resulted in an assessment framework containing ten criteria with a high degree of support of the experts in both countries. These criteria represented the

features that were to be measured, and specified the types of clinical encounters that needed to be simulated in order to accomplish this. The ten criteria and their full definitions are further described in appendix 1.

3.1.3.1.2 Global ratings

The ten criteria resulting from the Delphi-study were not further specified in sub-criteria, since the reliability of global ratings in (medical) examinations (with SPs) is known to be on par with more specific behaviorally anchored ratings (Cunnington, Neville & Norman, 1997; Gipps, 1994; Goaverts et al., 2007; Hawkins & Holmboe, 2008; Regehr, MacRay, Reznick & Szalay, 1998; Solomon, Szauter, Rosebraugh & Callaway, 2000). Moreover, global ratings 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 for unfamiliar situations) (Hawkins et al., 2009). Detailed (dichotomous) checklists (e.g. makes eye contact, introduces self) for such broad constructs often fail to validly capture the essential, often difficult to quantify, elements of expert behavior (Hawkins & Holmboe, 2008). The exclusion of checklists also increased the manageability of the UHTRUSTs' already demanding and complex scoring procedure; nurses for example were responsible for the fulfillment of *four* ten-page analytic scoring rubrics.

3.1.3.2 Holistic ratings: the entrustment decisions

3.1.3.2.1 The identification of entrustable professional activities (EPAs)

Besides the quality decisions concerning the respondents' performance on the separate criteria, judgment was assigned to the whole performance in the form of a series of entrustment decisions. At the end of the day the assessing *physicians* were asked to indicate for each observed respondent to what extent they would trust this person with a new critical clinical activity. They were asked to imagine that they bared final responsibility for the patients involved and that these patients were close relatives, personal friends, or publicly well known persons.

In order to do so, ten 'entrustable' professional activities (EPAs) were identified. An EPA is a professional activity that requires several specific competences and that should only be entrusted upon a competent enough professional (Ten Cate, 2005; Ten Cate & Scheele, 2007). The EPA scenarios were designed by a Dutch and German professor of medical education. The first mentioned introduced the EPA concept (Sterkenburg et al., 2010). The last mentioned also was a specialist in internal medicine. These two experts incorporated several of the identified features underlying the construct of interest in each EPA scenario. For example, one EPA scenario (EPA 10) comprised interaction with a medical consultant. A high entrustment score on this item was expected to correlate with a high score on for example criterion 2 (knowing and maintaining own personal bounds and possibilities), criterion 4 (teamwork and collegiality), and criterion 6 (verbal communication with colleagues and supervisors).

Pearson Correlations were calculated in order to verify these relationships. Table 3 shows which analytic scores on the scoring criteria were most predictive for the holistic scores on the EPA scenarios (or vice versa). Criteria 6, 8 and 10 significantly correlated with

all ten EPA scenarios: $p \leq .05$. Criterion 5 did not have any significant correlations with the EPA scenarios: $p > .05$. None of the scores on the EPA scenarios had high correlations with *all* analytic scores. Not all expected relationships were found; EPA 10 for example did not significantly correlate with criterion 4.

3.1.3.3 Construct validity

The heterogeneous test population of the UHTRUST pilot version made it possible to evaluate the distinctive character of the assessment; the power of the test to detect differences in ability. The question was whether or not the pilot assessment could discriminate between expert and novice medical practitioners. The criteria and scenarios were expected to closely match the proficiency level of the attending final year students and residents. Were junior medical students were expected to score below the average of the pilot test population, experienced GPs were expected to perform above.

When the test scores were analyzed, it was found that the GPs' *analytic* mean score was not the highest but the *lowest* of all participating groups (see Table 4). Dutch assessors stated that often it was not difficult to see which respondents were the oldest and most experienced, and that they had found it hard to ignore this during the rating process. They indicated that it was possible that expectations had biased their ratings; were a performance of a supposed student was marked as 'acceptable', the same performance of a supposed GP was marked as 'poor'. Therefore, it was decided that the GPs' test scores would be excluded from

Table 3. Pearson Correlations for All Respondents Scores on the Analytic Scoring Criteria, and the Holistic EPA Scenarios

	EPA1	EPA2	EPA3	EPA4	EPA5	EPA6	EPA7	EPA8	EPA9	EPA10
Criterion 1	.55*	.42	.41	.36	.31	.63**	.45	.48*	.49*	.56*
Criterion 2	.75**	.50	.50*	.41	.57**	.77**	.79**	.58**	.63**	.58**
Criterion 3	.64**	.45	.56*	.59**	.55*	.62**	.48*	.66**	.67**	.69**
Criterion 4	.48*	.54*	.50*	.30	.52*	.45	.71**	.31	.33	.41
Criterion 5	.23	.50	.25	.23	.26	.26	.34	.18	.22	.43
Criterion 6	.71**	.60**	.51*	.59**	.61**	.74**	.58**	.67**	.78**	.73**
Criterion 7	.53*	.69**	.59*	.60**	.53*	.56*	.58*	.51	.50	.80**
Criterion 8	.65**	.74**	.59**	.68**	.72**	.70**	.76**	.64**	.57**	.66**
Criterion 9	.59*	.76**	.69**	.61*	.51*	.39	.33	.59**	.46	.68**
Criterion 10	.73**	.65**	.65**	.66**	.73**	.64**	.66**	.65**	.59**	.59**

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 4. Analytic and Holistic Mean Scores of the Participating Groups

Group	N	Analytic mean	SD	Holistic mean	SD
Junior students	5	2.75	.56	2.81	1.02
Last year students	7	3.06	.62	2.89	.89
Residents	6	3.46	.50	4.27	.85
GPs	2	2.70	.90	3.65	2.26

further analyses. However, an one-way ANOVA showed no significant effect between the *analytic* mean scores of the remaining groups either: $F(2, 15) = 2.22$, $p = .14$, $\eta^2 = .23$. With the GPs excluded, a significant effect was found between medical experience and the *holistic* mean scores: $F(2, 15) = 4.78$, $p = .03$, $\eta^2 = .39$. Here, the Bonferroni post hoc test revealed significant differences between last year medical students (mean = 2.89) and residents (mean = 4.27): $p = .05$. A significant difference between the holistic mean scores of the junior medical students (mean = 2.81) and the residents (mean = 4.27) was nearly detected: $p = .06$.

In order to collect more statistical evidence on the construct validity of the pilot assessment it was verified whether the analytic and holistic scores were driven by the same underlying construct i.e. the ability to take responsibility for unfamiliar clinical situations. A principal component analysis with oblique rotation (direct oblimin) indicated that for the *analytic* scores a component that explained 60.1% of the variance existed. The pattern matrix showed that all individual factor loadings were $> .4$, except the loading of criterion 4 (loading = .11). A second component with an eigenvalue over Kaiser's criterion of 1 was found. Criteria 2, 4 and 5 had higher individual loadings on this component than on the first. However, the scree plot was slightly ambiguous, and because no clear interpretation of this second component was found it was declined. Considering the careful design of the analytic scoring criteria (see above) it was decided that all ten criteria belonged in the first, approved component. Together, the analytic scoring criteria formed a reliable scale: the Cronbach's alpha was .91. A second principal component analysis with oblique rotation (direct oblimin) showed that the *holistic* scores also reflected a single construct. Here, one component was found that explained 79.6% of the variance. No other components with an eigenvalues ≥ 1 were found, and the scree plot was unambiguous. The Cronbach's alpha of the ten holistic scenarios was .97.

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

3.1.4.1 Assessors and assessor preparation

For PAs, the quality of the assessment as a whole is considered with respect to the ability of the assessors to use the scoring criteria and rubric to reach a technically and professionally defensible conclusion (Dwyer, 1995). Therefore, assessors should be selected with care and they should develop this ability during assessor training (Bakker, 2008; Hawkins et al., 2009; Kane, 2006).

3.1.4.1.1 Assessor selection and background

It is important that an assessor has the expertise to make profound judgments about a respondents' ability in the construct of interest. Therefore, the nurses and physicians of the UMCU and UKE that participated as assessors in the pilot assessment all were selected because of their active clinical- and, in most cases, teaching experience.

There were slight differences between the demographics (e.g. age, working experience, gender) of the Dutch and German assessors (see method), but no significant differences between the two groups existed. However, excessive differences between individual assessors did exist. For example, were the physician with the least amount of working experience had six years of clinical experience, the physician with the largest amount

of working experience had 40 years of clinical experience. Also, some physicians had no teaching experience at all, were others had up to 22 years. The interchangeability of the individual assessing physicians of the UHTRUST pilot version can be questioned, because the excessive differences between individual assessors were most likely accompanied with highly different frames of reference (Bakker, 2008). However, in real-life these diverse individuals also tutor medical students, estimate their abilities, and the amount of responsibility they can handle. All assessors are qualified to do so. These differences can therefore also be seen as a contribution to the assessments' authenticity.

All assessors of the pilot assessment participated voluntarily. An open question concerning the motivation for participation on the evaluation forms showed that Dutch assessors mostly signed up because they wanted to extend their experiences as assessors, make a contribution to medical education, and because they were curious about the research questions concerning the curricula. Most German assessors indicated that they were curious about this new, more practice oriented way of assessing students. The motivation of assessors is important, given that high motivation make a contribution to the quality of the rating outcomes (Govaerts et al., 2007).

3.1.4.1.2 Assessor training

For the UHTRUST pilot version the scoring procedure was carefully implemented. An assessor training was set up to prepare the assessors for judging and scoring the performances of the respondents of the pilot assessment. Again, language barriers and physical distance made it impossible to work with both the German and Dutch assessors on the same occasion and location. Therefore, also the assessor training was standardized, so both training sessions resembled one another as much as possible. A half-day training session was developed in which the assessors learned to apply the detailed *scoring procedure* (see method) in a systematic and consistent way. During the training the assessment purpose and context, the ten scoring criteria, the ten EPA scenarios, the six SP scenarios, and the scoring rubrics were discussed. The scoring method was practiced step by step, and the assessors received feedback. During the training the assessors were corrected when they deviated from the scoring procedure.

Even though the experts of the Delphi study already reached consensus about the content and formulation of the scoring criteria, the criteria statements had to be operationalised by the assessors themselves to be understood in concrete and situation specific terms (Knight, 2002). During the training session the assessors had to develop common conceptualizations of what constitutes competent behavior. This was important, given the fact that all analytic criteria of the UHTRUST pilot version refer to abstract qualities that assessors need to infer from the respondent performance. What exactly is '*active professional development*' (criterion 3) or '*coping with mistakes*' (criterion 9)? Therefore, each of the ten criteria was illustrated with a series of examples of adequate and inadequate behaviors. These examples and counterexamples of proficient behavior were provided on video or by means of verbal cases, and were used to evoke discussion among the experts. In this discussion assessors challenged one another's interpretations. In this way the acceptability and tenability of the interpretations were critically checked, and the influence of personal characteristics on judgments were reduced as much as possible. The assessors reached consensus about

performance standards and rating scale anchors for each of the scoring criteria. These behavioral anchors concretize a 'Poor', 'Acceptable', or 'Excellent' performance and assist the assessors in making relevant and consistent considerations during the scoring process. According to Knight (2002) there always will be a certain degree of ambiguity about the meaning of criteria and interpretation of the standards. On this area, assessors will never be completely interchangeable. He claims that attempts to reach full shared understandings will threaten academic freedom. It is important to put trust in the judgment of an expert and not to quell creativity (Ten Cate, 2006; Gipps, 1994).

Another important goal of the training was to enhance the rater accuracy by making the assessors aware of rating errors; they were urged to correct those errors immediately if they occurred. Special attention was given to errors concerning range restriction (central tendency), rater leniency/severity, and halo-effect. Although both systematic (e.g. bias) and random (e.g. memory) sources of rater error cannot be eliminated entirely, they can be effectively controlled, thus alleviating some potential threats to the validity of assessment scores (Hawkins & Holmboe, 2008).

3.1.4.2 *Quality control data (validity)*

The scoring criteria and rubrics, the detailed scoring procedure, and the assessor selection procedure and training were all designed and implemented to enhance more objective and reliable scoring. However, they do not *guarantee* high quality assessment processes (Nijveldt, 2007). That is why the effects of these measures were statistically examined.

3.1.4.2.1 The assessors' use of the analytic scoring criteria and rubric

During the development of the pilot assessment test developers indicated that assessors would probably be unable to make valid judgments about criterion 5 (active listening to patients) and criterion 7 (empathy and openness, also to patients), since none of the assessors were present during the clinical encounters and none had spoken to the SPs afterwards. Because the selection of these two criteria by the experts of the Delphi study indicated that social and professional interaction with the patients was considered an important part of the construct, the SPs of the pilot assessment were asked to fill in a translated version of the consultation and relational empathy (CARE) measure directly after each encounter with a respondent. The CARE measure is a validated questionnaire for measuring patients' perceptions of relational empathy in consultation, developed by Mercer, Maxwell, Heaney and Watt (2004). The use of the instrument prevented that an information gap would arise when assessors indeed showed to be uncertain about their judgments of criteria 5 and 7.

After the pilot assessment the analytic scores were analyzed in order to verify these expectations. It was found that the item non-response was the highest for criteria 5 and 7. Despite of the fact that all respondents were evaluated by multiple assessors, six out of 20 respondents did *not* receive an analytic score for criteria 5 and 7 from *any* of their assessors (see Table 5, first column). Furthermore, analysis of the 3-point certainty scales showed that on average assessors also were the most uncertain about their judgments of these two criteria: the mean score of criterion 5 was 1.6, and the mean score for criterion 7 was 1.4 (see Table 5, second column). However, it should be noted that most assessors had not consequently used these certainty scales (see Table 5, third column). Again criteria 5 and 7 were most often

neglected; out of a maximum of 73 times, the certainty scale for criterion 5 was filled in on 17 occasions, and the certainty scale of criterion 7 was filled in on 16 occasions.

The missing data on the analytic scoring rubrics were not desirable. For one thing, they affected the total analytic mean scores of the respondents, and hence the results of for example the in the previous assumption mentioned ANOVA. For another, they were a sign that the detailed scoring procedure was not completely followed by all assessors. Despite the fact that on the evaluation forms assessors ($N = 19$) indicated that they thought the assessor training was adequate (68% of the assessors gave this item score 4, and 32% of the assessors gave this item a score 3), the training had not prevented that some assessors deviated from the prescribed scoring procedure. If they had followed the procedure, they would not have omitted to fill in an analytic score but they would have indicated their uncertainty about this score on the certainty scale. Time pressure also cannot explain the large amount of missing analytic scores; on the evaluation form assessors ($N = 17$) were asked if they had received a sufficient amount of time to score the observed performances on the scoring rubrics. Here, the mean score was 3.65 ($SD = 0.69$), and on a 5-point scale 65% of the assessors scored this item ≥ 4 . All in all, the current data seems to indicate that the concerns of the test developer with regard to scoring criteria 5 and 7 were grounded. When it is decided to remove these two scoring criteria from the analytic scoring rubric, the Cronbach's alpha for the eight remaining criteria still form an acceptable scale of .88.

3.1.4.2.2 The assessors' use of the holistic scoring rubric

Analysis of the holistic scores showed that physicians filled in the holistic scoring rubrics very accurately; all respondents had received holistic scores from all their assessors on all ten EPA scenarios. Two respondents were an exception; one of them missed a holistic score on EPA 7, the other missed a holistic score on EPA 8. Both of these respondents did receive holistic scores on these EPAs from 2 other physicians, so the missing data only had a small impact on the total holistic mean scores.

Table 5. Missing Analytic Scores and the Results and Use of the Certainty Scale on the Analytic Scoring Rubric

	Number of respondents ($N = 20$) without analytic scores	Mean score on the certainty scale*	Number of occasions that assessors filled in the certainty scale**
Criterion 1	1	2.3	47
Criterion 2	0	2.3	55
Criterion 3	1	2.0	37
Criterion 4	2	1.9	34
Criterion 5	6	1.6	17
Criterion 6	0	2.5	50
Criterion 7	6	1.4	16
Criterion 8	0	1.9	47
Criterion 9	4	1.8	32
Criterion 10	0	2.0	44

*All mean scores had standard deviations < 1

**In theory the certainty scale could have been used 73 times per criterion

Table 6. Agreement Between Assessor Groups on the Analytic Scores

Assessor group	N	Jury alpha	Scoring criteria the jury alpha was based on
1	4	.85	1, 2, 4, 6, 8, 10 (60%)
2	3	.91	1, 2, 3, 6, 8, 9, 10 (70%)
3	4	.89	1, 2, 4, 6, 8, 10 (60%)
4	3	.97	1, 2, 3, 6, 8, 9, 10 (70%)

3.1.4.3 Internal consistency (reliability)

Further analysis of the performances of the assessors was done by means of the calculation of jury alphas. The analytic scoring rubrics were filled in by 20 assessor groups (one assessor group per respondent). However, because of missing data, reliable calculations could only be made for 4 of these groups. EPAs were sometimes not filled in by an assessor group, but by an assessor pair. Here, the jury alphas were calculated for 6 assessor groups and 13 assessor pairs. The assessing physician of one respondent did not have a partner, so no jury alpha could be calculated (see method).

For the *analytic* scores jury alphas were calculated for assessor groups who had *no* missing data on at least 50% of the criteria (see Table 6). All Dutch assessor groups (N = 12) and four of the German assessor groups did not meet this criterion. The jury alphas for the remaining four German assessor groups were satisfactory, ranging from .85 to .97. They were based on the scorings of 6 or 7 analytic scoring criteria.

The jury alphas for the holistic mean scores were satisfactory for 10 assessor pairs and 3 assessor groups, ranging from .74 to .96. The jury alphas were low or even negative for 3 assessor pairs and 3 assessor groups, ranging from -17.78 to -.56.

To check the agreement between the total analytic mean scores and the total holistic mean scores, a paired-samples t-test was conducted. The results showed that the total mean of the analytic scores (3.07; SD .62) was not significant higher or lower than the total mean of the holistic scores (3.36; SD 1.17).

3.2 Inference 2: generalization

“If test scores are not reproducible, there is no basis for making an interpretation” (Kane, 2006).

In this second part of the argument the generalizability of the observed scores to the broader test domain is discussed. Table 7 provides a summary of the generalization inference.

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

3.2.1.1 Controlling sources of random error; obtaining stable test scores

Any facet that is allowed to vary in the universe of generalization (e.g. tasks, occasions, assessors) and is sampled by the measurement procedure contributes to the random error (Kane, 2006; Lane & Stone, 2006). Therefore, in the method and the first part of the argument

Table 7: Summary of the Generalization Inference

Assumptions underlying the generalization inference	Warrants licensing the assumptions
2.1: The scores were stable and random error due to different occasions, raters and tasks was controlled	<p>The <i>standardization measures</i> described in the first inference controlled the random error caused by administration occasion, rater, and tasks.</p> <p><i>Multiple assessors</i> per respondent were used to reduce the influence of personal biases of the individual assessors</p> <p>Respondents were confronted with <i>multiple cases</i>. This reduced the variance caused by tasks specificity, and provided the respondents with the opportunity to demonstrate their skills on multiple occasions.</p> <p>A generalization study is yet to be conducted.</p>
2.2: The sample of observations was representative of the universe of generalization	<p>The time consuming nature of the assessment tasks only made it possible to sample a <i>relatively small number</i> of assessment tasks.</p> <p>In order to compensate this deficiency, serious <i>effort</i> was made to draw a representative sample from the universe of generalization.</p> <p><i>Experts</i> were consulted during the task development and evaluation. They planed the content of the assessment in a <i>blueprint</i>, in order to make sure that the task sample could not be completed without the use of the skills of interest.</p> <p>The tasks and the task sample will be reviewed and modified after the pilot assessment.</p> <p>The representativeness of the task sample will always be open to debate.</p>

aspects of the measurement procedure that were considered fixed were described in detail (e.g. the assessment and scoring conditions, the assessor and SP training, the assessment content, and the evaluation tools). These standardization measures were implemented to control the random error caused by three variables; the administration occasion, the assessors, and the tasks. Besides these standardization measures, other countermeasures were implemented to prevent problems with reliability caused by these variables.

First, the stability of the test scores was enhanced by the use of multiple assessors per respondent. This reduced the influence of personal biases of the individual assessors (Kane, 2006). After all, even trained assessors may introduce potential sources of individuality into

the assessment process (Nijveldt, 2006). Second, the use of multiple encounters compensated the psychometric limitations inherent to a single encounter. On the assessment-day respondents were confronted with six cases, designed to sample skills more broadly over the course of the assessment. This reduced the influence of potential error caused by for example differences in SP case portrayal, differences in complexity of the individual cases, and variance caused by task-specificity (Gipps, 1994; Hawkins & Holmboe, 2008; Lane & Stone, 2006). The use of multiple cases also enhanced accuracy of the test scores, because the assessors were given the chance to judge the respondents clinical skills based on their performance on multiple occasions. Or, from a different perspective, the use of multiple cases provided the respondents with several opportunities to demonstrate their true clinical competence.

3.2.1.2 Generalization study

At the moment, not all data that was generated during the pilot assessment has been used in its full potential. A generalization study is yet to be conducted. By means of this study, the overall variance of the test will be divided into multiple sources of error. Analysis-of-variance will make it possible to estimate the role of errors contributed by for example variability in assessment tasks and assessors. The generalization study can make apparent if inferences can be drawn that extend beyond the participants of the pilot assessment.

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

3.2.2.1 Content representativeness

Any assessment is but a sample of the tasks that could be presented to respondents (Kane, 2006). As mentioned in the theoretical framework of this study, the selection of *representative samples of assessment tasks* is an important issue in PA (Bakker, 2008). This also was the case with the UHTRUST pilot version; the authentic character of the pilot assessment and the time-consuming nature of its tasks only made it feasible to confront respondents with six unfamiliar encounters. This is not an excessive number, given that previous research indicated that reliable assessment of clinical skills often require six to ten cases (Hawkins & Holmboe, 2008). Therefore, following advice of Kane (2004), serious effort has been made to draw a representative sample from the universe of generalization. The efforts taken are described below.

Firstly, the document-, literature- and Delphi study were used to identify the critical features that were considered most important to assess (Appendix 1). Logically, these were the skills that had to become apparent in the performance elicited by the pilot assessment. Secondly, to make sure that the tasks of the pilot assessment could not be completed without equitably using these skills, ten medical experts were consulted during tasks development and evaluation. This group of experts consisted of eight physicians and the two aforementioned professors of medical education. These experts were asked for their clinical opinions regarding how a case should play out, and agreed that the tasks included in the pilot assessment portrayed a sufficiently broad content. Also, they developed a blueprint to plan the content of the test in such a way that it was considered representative for the test domain.

Thirdly, the tasks within the pilot assessment comprised more than just the six encounters. The respondents' ability to take responsibility for unfamiliar situations was also judged by the way they handled the disruptions in the second phase of the assessment, and the way they explained and defended their considerations and actions during the reporting phase in the third phase of the assessment. These additional tasks were also thoroughly planned and discussed by the experts and test developers.

In defence of these efforts, Eisner (1991) pointed out that sampling assumptions are rarely satisfied, and therefore "inferences are made to larger populations, not because of impeccable statistical logic, but because it makes good sense to do so" (Eisner, 1991, p. 203). Also, it should be remembered that the data of the pilot assessment will be used to further improve the tasks and the task sample of the UHTRUST.

3.3 Inference 3: extrapolation

A highly reliable score is of no value if it measures the wrong characteristics (Hawkins & Holmboe, 2008).

In this third part of the argument the extent to which the performances on the sampled test tasks from the universe of generalization can be extrapolated to performance in practice is discussed. Table 8 provides a summary of the extrapolation inference.

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

3.3.1.1 High fidelity simulation

Serious effort was made to achieve a high level of physical and psychological fidelity. First, to obtain a high level of physical fidelity, all rooms were converted into consulting rooms and doctor offices. Also, the choice to work with SPs (instead of for example written patient scenarios) contributed to the level of physical and psychological realism; in performing the simulation, the SP does not only presents the gestalt and history of the patient being simulated, "but the body language, the physical findings, and the emotional and personality characteristics as well" (Cleland, Abe & Rethans, 2009, p. 478). The psychological fidelity was further enhanced by the tasks, and the modes of presentation. The tasks were designed in such a way that they could also occur on a real clinical ward, and respondents were free to react to these tasks as they would have done in real life. Lane and Stone (2006) stated that these kinds of high-fidelity tasks can easily be translated to expected performance in the real-world. Last, the act of observing could have interfered with the level of authenticity, was it not for the fact that respondents were never observed by their assessors at unrealistic moments. For example, none of the assessors were present during the clinical encounters in the first phase of the pilot assessment, and during the period of study and thought in the second phase of the pilot assessment only the nurses were present on the simulated ward. The assessing physicians could only observe the respondents during their scheduled visit to the simulated ward, or when a respondent called his supervisor for consultation.

Table 8: Summary of the Extrapolation Inference

Assumptions underlying the extrapolation inference	Warrants licensing the assumptions
3.1: The universe score was related to the level of skill of the graduate in the target domain	<p>The <i>authentic character</i> of the pilot assessment makes the argument for extrapolation plausible.</p> <p>When a comprehensive construct is measured, the practical <i>limits of assessment</i> must be accepted.</p> <p>The pilot assessment provided <i>negative evidence</i> on the respondents' true ability to cope with unfamiliar clinical situations.</p>
3.2: There were no systematic errors that were likely to undermine the extrapolation	<p>The standardized assessment conditions and the use of standardized patients brought about an <i>artificial</i> aspect to the pilot assessment.</p> <p>Sources of irrelevant variance caused by <i>systematic differences between SPs (and real patients), time pressure, and systematic differences between respondents</i> were identified and controlled.</p>

The effect of these efforts was evaluated afterwards; on the evaluation forms both respondents, and assessors were asked if the pilot assessment had a high level of authenticity. The respondents (N = 18) mean score on this item was 4.3 (SD = .66), the assessors (N = 20) mean score was 3.9 (SD = .55). On a 5-point scale 84% of the respondents and assessors (N = 38) scored this item ≥ 4 .

3.3.1.2 Negative evidence

As mentioned the above, during the development stage of the pilot assessment document- and literature studies were conducted, and experts were consulted to make sure that the assessed features were considered critical to successful practice outcome. The success of these efforts was checked among respondents and assessors of the pilot assessment. On the evaluation form they were asked if the assessed skills were relevant to the practice domain. The means score from the respondents (N = 20) on this item was 4.40 (SD = 0.59), and the mean score from the assessors (N = 19) was 4.15 (SD = 1.3). On a 5-point scale 87% of the respondents and assessors (N = 39) scored this item ≥ 4 .

The basic assumption here is that that the activities in the test domain are necessary for effective dealing with unfamiliar clinical situations in the practice domain. This assumption makes it reasonable to assume that respondents of the pilot assessment who were successful on the test also would be successful in the real unfamiliar situations. However, this cannot be taken as absolute evidence. Using an assessment task that closely approximates the practice setting has the potential to limit the effects of construct-irrelevant variance and construct underrepresentation, but this similarity does not ensure that the score appropriately represents

the proficiency of interest (Hawkins & Holmboe, 2008). For example, when a respondent demonstrates that he is *capable* to take responsibility for unfamiliar situations during the test, this is no guarantee that this behaviour is (always) *manifested* in real clinical setting. Also, it is no guarantee that the respondent is able to deal with unfamiliar clinical situations that fall outside the used task sample. This will be the case when a respondent *lacks other essential skills or features* (e.g. motor coordination or character) that were not included in the test. The respondents of the pilot assessment for instance did not have to perform any physical examinations in the first phase of the pilot assessment. Instead, the necessary information was handed to them on paper. Limits to assessment have to be accepted (Knight, 2004), but the fact remains that in real life the ability to perform accurate physical examinations *is* related to the ability to take responsibility for unfamiliar clinical situations. Kane (2004) therefore stated that for most PAs this assumption tends to be stronger on the negative side than it is on the positive side; even though not every aspect of the construct can be measured, it is reasonable to assume that a respondent that showed serious *deficiencies* in the test domain would also show *deficiencies* in the practice domain. For the UHTRUST pilot version this means that the test scores should be seen as *negative evidence* for the respondents' abilities in real life.

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

3.3.2.1 Controlling sources of systematic error; obtaining meaningful test scores

Despite of its authentic character, the purpose of the UHTRUST project asked for standardization of aspects that are not also fixed on an actual ward or clinic. As mentioned above, among others the assessment and scoring conditions were fixed for all respondents. Some of these standardization measures brought about an artificial aspect to the pilot assessment, and resulted in sources of systematic error that had to be identified and controlled.

3.3.2.1.1 The performances of the SPs

First, also trained SPs can be a source of irrelevant variance. Hawkins and Holmboe (2008) stated that even though little research on the subject exists, it is inevitable that differences between SPs and real patients exist. For this reason, the acceptability of the performance of the SPs during the pilot assessment was checked on the evaluation form. Respondents (N = 20) gave the plausibility of the performances of the SPs an average score of 4.55 (SD = 0.58). On a 5-point scale 95% of the respondents scored this item ≥ 4 (score 4 = 25%, score 5 = 70%).

Because the training of the German and Dutch SPs took place on separate places and occasions, the possibility of systematic differences between the performances of these two groups was taken into account. Also, in Hamburg patient scenarios were not acted out by one, but two different actors. Consequently, there was a risk of systematic differences between these two SPs. Potential differences in case portrayal can be caused by different personal experiences, and culture (Hawkins & Holmboe, 2008). However, because of the high contentment of the respondents with the performances of the SPs, the quality of their performances was not further investigated.

3.3.2.1.2 Time pressure

Second, time pressure can yield in invalid measures of proficiency when it contributes to construct-irrelevant variance (Hawkins & Holmboe, 2008). The medical experts that were involved in the development process of the assessment content indicated that the time allotted for the respondents to complete the individual phases of the pilot assessment was short, but realistic. However, they also stated that the actual time necessary for each phase could only be determined after the scenarios had been acted out during the pilot assessment.

On the evaluation form respondents (N = 20) were asked if they had received enough time to complete the tasks in the individual phases of the pilot assessment (four items). In table 8 the results are presented. On average respondents were quite satisfied with the amount of time they had received for various phases of the assessment: the mean scores on all four items were > 3. Assessors (N = 15) also were asked if they thought the respondents had received a sufficient amount of time to complete their tasks in the individual phases of the pilot assessment. Their mean scores on the four items were somewhat lower (see Table 9). It is possible that the assessors were more critical on the subject because of their rich concepts of what should have been done in the available amount of time (Woolfolk, Hughes & Walkup, 2008). Respondents on the other hand might be prematurely satisfied. Both for the respondents and assessors, scores on these four items were fairly distributed. It is not clear why opinions within a group contrasted.

During evaluation of the pilot assessment various respondents and assessors indicated that time pressure is often present in practice settings, and therefore should not be seen as irrelevant variance.

3.3.2.1.3 Trait implications

Third, following Kane (2006), it is presumed that the performances of the respondents during the pilot assessment are also influenced by other traits. In the case of the UHTRUST in particular the impact of culture, personality and medical knowledge are expected to explain some variance. Systematic differences between respondents on these traits can become a source of irrelevant variance when they are not accounted for. For example, despite of an *excellent* ability to cope with unfamiliar clinical situations, introvert respondents might not be perceived as ‘warm and friendly’ (a criterion on the CARE measure) by their SPs, or as ‘empathetic and open’ (criterion 7) by their assessors. The respondents’ introvert personality then becomes a source of irrelevant variance, because it negatively affect the respondents’ test scores. For the pilot assessment it was not feasible to take such differences into account.

Table 9. Evaluation of the Available Amount of Time Per Assessment Phase

<i>There was a sufficient amount of time for...</i>	Mean test respondents	SD	Mean assessors	SD
...the clinical encounters, phase 1	4.15	0.58	3.30	0.41
...the period of study and thought, phase 2	3.05	0.97	2.70	0.86
...the additional tasks, phase 2	3.15	1.03	3.05	0.93
...the reporting on the management proposals, phase 3	3.40	0.97	3.25	0.96

3.4 Inference 4: interpretation

Now that the extent to which the test scores can be generalized to the test domain and extrapolated to the practice domain is discussed, the last question that remains is whether or not the test scores can be used and interpreted as intended. This question is discussed in this final inference. Table 10 provides a summary of the interpretation inference.

3.4.1 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, the inferences and assumptions that underlie an argument for validity were made explicit in a theoretical framework in advance (Table 1). All available backing was used to support the identified assumptions, but, as mentioned earlier, at the moment not all data that was generated during the pilot assessment has been used in its full potential. It is recognized that the accuracy and plausibility of the backing that was presented in the generalization inference increases when the data of the generalization study is elaborated. This data will provide insight in the strength of the claims that were made in this part of the argument. The argumentation in the extrapolation inference will gain in strength when the interrelated traits are accounted for. Moreover, it remained disputable whether or not more validity evidence should have been gathered. Interviews with assessors for example could have thrown light on the quality of their cognitive processes and the effects of the assessor training. Results could have been included as backing in the argument for validity and could have strengthened the argumentation in the first inference. The possibilities for data collection were almost endless, but resources were not. Therefore, the main focus during the pilot study was put on testing the most prominent aspects of the PA: the planning, the standardization measures, the examination content, the assessor training, the scoring criteria, the EPA scenarios, the scoring rubrics, the scoring procedure, the authenticity level, and the investigation of potential threats to validity (Hawkins & Holmboe, 2008).

Table 10. Summary of the Interpretation Inference

Assumptions underlying the interpretation inference	Warrants licensing the assumptions
4.1: All assumptions were defensible with accurate and plausible evidence	<p>Most assumptions were defensible with accurate (and often parallel lines of) backing.</p> <p>Some assumptions (and hence the overall validity) will become more plausible when more backing is enclosed.</p>
4.2: The test scores can be used for the intended purposes	<p>The test scores of the pilot assessment cannot be used to answer the research question of the UHTRUST project.</p> <p>The pilot assessment did give an indication of the tests' validity in its current form.</p>

3.4.2 Assumption 4.2: The test scores can be used for the intended purposes

The UHTRUST is being developed to answer the question which of the aforementioned curricula better prepares medical graduates to take responsibility for unfamiliar clinical situations. The test scores of the pilot assessment cannot be used for this research purpose, simply because the sample of test respondents was not derived from the population of interest. However, the pilot assessment did give an indication of the tests' validity in its current form and information on how to improve its validity; based on the pilots' information the UHTRUSTs' planning, examination content, scoring rubrics, scoring procedure, and assessor training will be modified. Discussion of these changes is beyond the scope of this article.

4. Conclusion and discussion

The construction of an argument for validity is an iterative process which should lead to continued improvement in the quality and defensibility of an assessment (Kane, 2006). However, this iterative nature of the validation (and test development) process should not discourage a critical discussion of the current results; so, what is the validity of the UHTRUST pilot version?

4.1 The scoring inference

4.1.1 The scoring conditions

The test scores of the pilot assessment were acquired under generally fair assessment conditions. Serious effort was put into the design of score equating strategies. Most of these strategies were well implemented (e.g. the standardized content), but some irregularities between the two registration occasions did occur. For one thing, German respondents had received more time to complete the first phase of the assessment. This gave them an unfair advantage over the Dutch respondents. Also, because an unequal amount of respondents and assessors volunteered for each registration occasion, not all respondents were evaluated by the same amount of assessors, and not all assessors were paired up. Consequently, the quality of the ratings could not always be analyzed, and some respondents were more likely to suffer from (undiscovered) personal biases of individual assessors than others (Kane, 2006). In future registration occasions, such violations of the *ceteris paribus* assumption will negatively affect the research purpose of the UHTRUST project; it will make the claim that differences between medical graduates were mainly caused by their curricular background disputable. The problems encountered can easily be solved by more accurate planning, and the recruitment of assessors who are willing to be on stand-by (Boursicot & Roberts, 2005).

4.1.2 The analytic scorings

The analytic scoring criteria that enabled the measurement of the ability to take responsibility for unfamiliar clinical situations were carefully designed by means of a Delphi study. A principal component analysis on the analytic scores confirmed that the analytic scoring criteria were driven by a single construct, and together the ten scoring criteria formed a highly

reliable scale. The analytic criteria were unable to discriminate between medical novices (the students) and experts (the residents and GPs). Especially striking was the low analytic mean score of the GPs. This low score was probably caused by biased ratings; assessors had paid attention to irrelevant features. They had let respondents appearances guide their expectations and severity in scoring. The consequences of such violations should be more thoroughly addressed in future assessor training sessions, so assessors will become more aware of the quality of their own judgment process (Nijveldt, 2007). Furthermore, there was a large amount of item non-response on the analytic scoring rubrics. There are indications that the analytic scoring rubrics will be filled in more accurate when criteria 5 and 7 are abandoned and replaced by the CARE measure. This will make the complex scoring procedure of the UHTRUST more manageable. Also, when the analytic rubrics are filled in more accurately, it will become possible to make better estimations of the interrater reliability and calculate weighted analytic mean scores. The score on the certainty scale then determines the weight of an analytic score. It is expected that such weighted scores more accurately represent the true abilities of test respondents. In order to prevent problems with missing data in the future, the scoring procedure should be more carefully implemented, and the topic should be addressed during the assessor training. More thorough examination of the scoring rubrics before intake can also prevent large amounts of missing data in the future (Cohen & Wollack, 2006).

4.1.3 The holistic scorings

The holistic EPA scenarios were carefully designed by two professors of medical education and passed all the statistical tests for construct validity (although GPs had to be excluded as a group in the ANOVA). Unlike the analytic scoring rubrics, the holistic scoring rubrics were filled in very accurately. This made it possible to estimate the interrater reliability of all assessor pairs and groups. The jury alphas of most of these pairs and groups were satisfactory. Jury alphas were used to estimate the interrater reliability because they were the most feasible statistical measure. It should be noted that they only show how consistent assessors were in their own ratings. Jury alphas are also high when test scores of assessors vary systematically (Field, 2009). This sometimes resulted into a somewhat distorted picture of the interrater reliability. For example, one group of physicians had a negative jury alpha (-17.78), even though 40% of their judgments on the EPA scenarios were identical, whereas another group of physicians, who had more divergent opinions about the amount of trust a respondent deserved, had a very high jury alpha (.96). Therefore, Cohen's Kappa would have been a good supplement of the jury alphas.

4.2 The generalization inference

Experts developed the examination content of the pilot assessment de novo and indicated that the examination content was representative for the construct of interest; the tasks could not be resolved without the use of the critical features that were identified by the experts of the Delphi study. However, it remained difficult to substantiate this claim with indisputable evidence (see Eisner, 1991). In this case, this was also true because the construct was never clearly divined. What exactly is an '*unfamiliar clinical situation*'? Is it a patient with an unusual condition? A situation the doctor never experienced before? And does it include

working with unknown instruments or colleagues? Were the nurses mainly based their judgments on how a respondent handled the additional tasks during the second phase of the assessment, physicians based theirs on the respondents' (working towards the) management proposals. So, it is likely that an 'unfamiliar clinical situation' was something completely different to a nurse than to a physician. This ambiguity makes it hard to make claims about the representativeness of the used task sample, the presence or absence of construct irrelevant variance or underrepresentation, and to interpret a respondents' test score. In order to solve this problem and enhance the strength of the argumentation in the second part of the argument for validity, the exact meaning of the construct should be defined and discussed during the upcoming assessor training (Kane, 2006; Lane & Stone, 2006). When this will be done, it is important that the construct is not defined too narrowly; this will not match with the overall design of the PA.

4.3 The extrapolation inference

In general, PAs include a small sample of high-fidelity (but time consuming) performances. This makes generalization to a broadly defined universe of generalization often quite undependable, but the high authenticity level does empower extrapolation (Kane, 2006). This general rule also holds true for the UHTRUST pilot version; the task presentations and response formats were not (very) different from tasks in the target domain, and therefore only minor irrelevant method variance was detected. Particularly striking was the high respect of the respondents with regard to the performances of the SPs. Their excellent acting had contributed to the respondents' perceived level of realism. Whether or not time pressure was present and, if so, whether or not it biased extrapolation remains disputable

In the future, the strength of the extrapolation inference can be increased when external evidence is collected that verifies relationships between the observed scores and other scores or variables that are associated with the target domain (Lissitz & Samuelson, 2007).

4.4 The interpretation inference

Most assumptions were defensible with accurate and often parallel lines of backing, but missing data detracted the plausibility of others. This problem occurred because not all data that was gathered during the pilot assessment is fully processed yet (e.g. the missing generalization study). On that matter, the completion of this study was premature, even for an initial argument for validity. Furthermore, available resources did not make it feasible to further investigate the most questionable assumption of every rater-scored assessment: the quality of the cognitive processes underlying raters' judgment (Dweyer, 1995; Govaerts et al., 2007; Lane & Stone, 2006; Moss, 1994; Sterkenburg et al., 2010). To enhance the plausibility of this assumption it is advisable to collect such data during (and/or after) the next assessment registration, for example by means of (retrospective) verbal protocols or interviews with assessors. Bakker (2008), Nijveldt (2007), and Van der Schaaf, Stokking, & Verloop, (2005) are studies that can serve as an example.

When in future registration occasions of the UHTRUST the population of interest is used, it should be remembered that test scores should be interpreted as negative evidence for the respondents' true abilities.

4.5 Closing remarks

The complex context and design of the UHTRUST pilot version made it hard to write a short and robust argument for validity. The elaboration of the theoretical framework for an argument for validity (Table 1) was the heart of this study but the true argument for validity starts in the introduction (where the assessments' value, purpose, and context are defined) and only ends here, in the discussion (where the results are further discussed and recommendations are made). It is up to the reader to critically evaluate the overall clarity, coherence, and plausibility of the argument (Kane, 2006).

Future research should examine whether the finished assessment product can be used for its proposed interpretation and use. In a follow-up study more attention could be paid to differences between the various rater types (supervisors, physicians, and nurses). During the UHTRUST the different rater types observe different aspects of the assessment day, and some types see more of the respondents than others. Therefore, significant differences between rater types on the certainty scale of the *analytic* scoring rubrics can for example give an indication of how much a clinical assessor should be able to see of a student or trainee in order to make dependable judgments of *discrete* constructs (e.g. communication, or empathy). Also, following Sterkenburg and colleagues (2010), medical education could benefit from future research that address the question why physicians had put more *trust* into the abilities of some respondents than others. Furthermore, to make validation a more accessible enterprise for educational measurement practitioners critical evaluation of the framework (Table 1) and methodology used in this study can be of great value for future validity studies.

5. References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bakker, M. (2008). *Design and evaluation of video portfolios. Reliability, generalizability, and validity of an authentic performance assessment for teachers*. Leiden: Mostert & Van Onderen!
- Bonnetain, E., Boucheix, J., & Hamet, M. (2010). Benefits of computer screen-based simulation in learning cardiac arrest procedures. *Medical education*, 44 (7), 716-723.
- Boursicot, K., & Roberts, T. (2005). How to set up an OSCE. *The Clinical Teacher*, 2 (1), 16-20.
- Cate, O. T. J., ten. (2005). Entrustability of professional activities and competency-based training. *Medical Education*, 39, 1176-1177.
- Cate, O. T. J., ten. (2006). Trust, competence, and supervisor's role in postgraduate training. *BMJ: British medical journal*, 333 (7571), 748-751.

- Cate, O. T. J., ten. (2007). Medical education in the Netherlands. *Medical Teacher*, 29, 752-757.
- Cate, O. T. J., ten., & Scheele, F. (2007). Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? *Academic Medicine*, 82 (6), 542-547.
- 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 (1), 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 (6), 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.
- Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1997). This risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Science Education*, 1, 227-233.
- 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 Academic Publishers.
- Eisner, E. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.
- Field, A. (2009). *Discovering Statistics Using SPSS*. Thousand Oaks, California: SAGE Publications Inc.
- Gipps, C. V. (1994). *Beyond Testing. Towards a Theory of Educational Assessment*. London: RoutledgeFalmer.
- Govaerts, M. J. B., Vleuten, C. P. M. van der., 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*, 12, 239-260.
- 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 (4), 348-361.
- Hawkins, R. E., & Holmboe, E. S. (2008). *Practical Guide to the Evaluation of Clinical Competence*. Philadelphia: Mosby Elsevier.
- Kane, M. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement: Interdisciplinary Research & Perspective*, 2 (3), 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.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18 (2), 5-17.
- Knight, P. T. (2002). The Achilles' Heel of Quality: the Assessment of Student Learning. *Quality in Higher Education*, 8, 107-115.
- 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.

- Lissitz, R.W., & Samuelsom, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36 (8), 437-448.
- 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 (6), 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.
- Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.
- Nijveldt, M. (2007). *Validity in teacher Assessment. An exploration of the judgement processes of assessors*. Enschede: Gildeprint.
- Regehr, G., MacRay, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessment on an OSCE-format examination. *Academic Medicine*, 73 (9), 993-997.
- Schaaf, M. F. van der., & Stokking, K. M. (2008). Developing and validating a design for teacher portfolio assessment. *Assessment & Evaluation in Higher Education*, 33 (3), 245-262.
- Schaaf, M. F. van der., Stokking, K. M., & Verloop, N. (2005) Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation*. 31, 27-55.
- Solomon, D. J., Szauter, K., Rosebraugh, C. J., & Callaway, M. R. (2000). Global Ratings of Student Performance in a Standardized Patient Examination: Is the Whole More than the Sum of the Parts? *Advances in health sciences education*, 5 (2), 131-140.
- Sterkenburg, A., Barach, P., Kalkman, C., Gielen, M., & Cate, O. T. J., ten. (2010). When do supervising physicians decide to entrust residents with unsupervised tasks? *Academic Medicine*, 85 (9), 1408-1417.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Vleuten, C. van der., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, 2 (2), 58-76.
- Wijnen-Meijer, M., Cate, O. T. J. ten., Schaaf, M. van der., & Borleffs, J. C. C. (2010). Vertical Integration in Medical School: Effect on the Transition to Postgraduate Training. *Medical Education*, 44, 272-279.
- Woolfolk, A., Hughes, M., & Walkup, V. (2008). *Psychology in Education*. Essex: Pearson Education Limited.

Appendix 1

The Ten Analytic Scoring Criteria of the UHTRUST Pilot Version and their Definitions.

Scoring criteria	Definition
1. Scientific and empirical grounded method of working	The physician uses evidence-based procedures whenever possible and relies on scientific knowledge. He searches actively and purposefully for evidence and consults high quality resources. He uses his scientific knowledge critically and carefully in his work.
2. Knowing and maintaining own personal bounds and possibilities	The physician knows the boundaries of his own ability and asks for help (timely) when needed. He reflects on himself and the situation.
3. Active professional development	The physician aims for quality and professional development by means of a critical attitude towards himself and his environment, study, self assessment, reflection, asking for feedback and setting and achieving learning goals. He reacts to criticism constructively and is aware of his own responsibility regarding his own abilities.
4. Teamwork and collegiality	The physician cooperates effectively and respectful in a (multidisciplinary) team, taking the views, knowledge and expertise of others into account.
5. Active listening to patients	The physician listens actively to patients and reacts (verbally and nonverbally) on the things he hears in a way that encourages the sharing of information (by the patients) and confirms his involvement with the patient. He shows attention to non-verbal signals coming from the patients.
6. Verbal communication with colleagues and supervisors	The physician gives structured, pithy and unambiguous verbal reports on his findings on a patient and his diagnostic and therapeutic policy. He asks relevant and purposeful questions.
7. Empathy and openness	The physician shows empathy, openness and susceptibility/accessibility in his contact with patients.
8. Responsibility	The physician takes responsibility and shows accountability for his work. He accepts liability for his work.
9. Coping with mistakes	The physician is aware of the fact that anyone can make and makes mistakes once in a while. He is approachable when someone points out his mistakes and reacts

adequately when he thinks that a colleague makes a mistake.

10. Safety and riskmanagement

The physician is alert and critical. He recognizes risks and responds to them timely. He aims at safety by the use of protocols where possible or the deliberate deviation of these protocols for the benefits of the patient. He reports irresponsible behavior.