Assessing clinical reasoning (ASCLIRE): Instrument development and validation

Article *in* Advances in Health Sciences Education · March 2015

DOI: 10.1007/s10459-015-9596-y · Source: PubMed

5 authors, including:

Olga Kunina-Habenicht
Technische Universität Dortmund
63 PUBLICATIONS **1,108** CITATIONS
SEE PROFILE

Wolf E Hautz
Inselspital, Universitätsspital Bern
118 PUBLICATIONS **801** CITATIONS
SEE PROFILE

Claudia Spies
Charité Universitätsmedizin Berlin
963 PUBLICATIONS **25,207** CITATIONS
SEE PROFILE

Adv in Health Sci Educ
DOI 10.1007/s10459-015-9596-y

Assessing clinical reasoning (ASCLIRE): Instrument development and validation

Olga Kunina-Habenicht · Wolf E. Hautz · Michel Knigge · Claudia Spies · Olaf Ahlers

Received: 15 August 2014 / Accepted: 19 February 2015

Abstract Clinical reasoning is an essential competency in medical education. This study aimed at developing and validating a test to assess diagnostic accuracy, collected information, and diagnostic decision time in clinical reasoning. A norm-referenced computer-based test for the assessment of clinical reasoning (ASCLIRE) was developed, integrating the entire clinical decision process. In a cross-sectional study participants were asked to

O. Kunina-Habenicht (✉)
Department of Educational Psychology, Institute of Psychology, Goethe-University Frankfurt, Theodor-W.-Adorno-Platz 6, 60629 Frankfurt am Main, Germany
e-mail: kunina@paed.psych.uni-frankfurt.de

W. E. Hautz · C. Spies · O. Ahlers
Department of Anesthesiology and Intensive Care Medicine CVK/CCM, Charité – Universitätsmedizin Berlin, 10117 Berlin, Germany
e-mail: wolf.hautz@insel.ch

C. Spies
e-mail: claudia.spies@charite.de

O. Ahlers
e-mail: olaf.ahlers@charite.de

W. E. Hautz
Universitäres Notfallzentrum, Inselspital Bern, 3010 Bern, Switzerland

M. Knigge
Department Educational Science, Institute of Educational Psychology, Franckeplatz 1, 06099 Halle (Saale), Germany
e-mail: michel.knigge@gmail.com

M. Knigge
Human Sciences Faculty, University of Potsdam, Karl-Liebknecht-Str. 24-25, Building 31, 14476 Potsdam, Germany

O. Ahlers
Department for Curriculum Management, Charité – Universitätsmedizin Berlin, 10117 Berlin, Germany

Published online: 01 March 2015

choose as many diagnostic measures as they deemed necessary to diagnose the underlying disease of six different cases with acute or sub-acute dyspnea and provide a diagnosis. 283 students and 20 content experts participated. In addition to diagnostic accuracy, respective decision time and number of used relevant diagnostic measures were documented as distinct performance indicators. The empirical structure of the test was investigated using a structural equation modeling approach. Experts showed higher accuracy rates and lower decision times than students. In a cross-sectional comparison, the diagnostic accuracy of students improved with the year of study. Wrong diagnoses provided by our sample were comparable to wrong diagnoses in practice. We found an excellent fit for a model with three latent factors—diagnostic accuracy, decision time, and choice of relevant diagnostic information—with diagnostic accuracy showing no significant correlation with decision time. ASCLIRE considers decision time as an important performance indicator beneath diagnostic accuracy and provides evidence that clinical reasoning is a complex ability comprising diagnostic accuracy, decision time, and choice of relevant diagnostic information as three partly correlated but still distinct aspects.

Keywords Clinical reasoning · Computer-based assessment · Diagnostic accuracy · Medical education · Validation · Decision speed

Abbreviations

CFI	Comparative fit index
CT	Computed tomography
df	Degrees of freedom
ECG	Electrocardiogram
MCQs	Multiple choice questions
OSCEs	Objective structure clinical examinations
RMSEA	Root mean square error of approximation
SEM	Structural equation model
WLSMV	Weighted least standardized means und variance

Background

The diagnostic process is one essential part of medical competencies (Epstein and Hundert 2002) and objectives related to its mastery are part of medical education standards (Hays 2014) and licensing procedures (Swing 2007). The diagnostic process is discussed under different terms (e.g. “clinical reasoning”, “medical problem solving”, “medical decision making”, or “medical judgment”) in the literature (for a review see Elstein 2009; Eva 2004; Norman 2005). There is broad agreement that ‘clinical reasoning, or one of its many synonyms <…> should be taught and tested’ (Norman 2005, p. 418).

Documentation and evaluation of clinical reasoning is essential to avoid wrong diagnoses and possibly resulting wrong treatments. Such wrong diagnoses often result from cognitive errors such as selective data perception, wrong data interpretation, missing relevant professional knowledge, or wrong data synthesis. Faulty diagnoses in turn substantially contribute to morbidity and suboptimal therapy (Gandhi et al. 2006; Graber et al. 2005; Kachalia et al. 2007; Newman-Toker and Pronovost 2009; Singh et al. 2007). Current estimates of diagnostic errors in medicine range from 2 to 5 % in visual specialties

like radiology or pathology to up to 15 % or more in the emergency room (Berner and Graber 2008).

Processes governing clinical reasoning and its relation to decision time

During the last four decades numerous models of clinical reasoning have been proposed (Charlin et al. 2012; Norman 2005). Most studies treat clinical reasoning as a behavior-oriented form of concrete reasoning within a given medical context (Norman 2005).

Currently, mostly two cognitive strategies have been distinguished: analytic processes, termed 'system 2 reasoning' and non-analytic processes, termed 'system 1 reasoning' (Eva 2004). System 1 reasoning is fast due to parallel information processing, unconscious, not accessible to meta-cognitive control and often perceived as pattern recognition, pattern discrimination, or heuristic reasoning (Croskerry and Norman 2008; Norman and Brooks 1997). System 2 reasoning in turn is slower due to serial information processing (Croskerry and Norman 2008) and accessible to intentional strategy selection.

Some authors described diagnostic errors as more strongly related to the fast system 1 reasoning than the analytic but rather slow system 2 reasoning (Croskerry 2009). This phenomenon has often been called "accuracy-speed- trade-off" implying that one person is either slow and accurate, or fast, but more likely to make mistakes (Wickelgren 1977). Several previous studies demonstrated contradicting results: Norman et al. (1989) found that correct diagnoses were associated with a decrease in response time, while errors were related to a substantial increase in response times. Recently, Sherbino et al. (2012) presented similar results for a representative sample using a computer-based tool for clinical multiple choice questions (MCQs). The authors investigated system 1 reasoning and reported a strong correlation of $r = -.54$ between accuracy and response time, meaning that 'increased accuracy was consistently associated with decreased time and, hence, greater speed' (p. 788). Consistent with this finding, Norman et al. (2014) have demonstrated that 'simply encouraging slowing down and increasing attention to analytical thinking was insufficient to increase diagnostic accuracy' (p. 277). Thus, evidence on the relationship between accuracy and decision speed seems to be inconsistent and was mostly studied using MCQ. Moreover, none of the studies, which have examined the total time for problem solving, considered the number of necessary diagnostic steps or the amount of used diagnostic information.

Role of expertise and the amount of used information

General psychological literature (Bohle Carbonell et al. 2014; Ericsson et al. 2006) and medical literature on expertise (Lesgold et al. 1988; Norman et al. 1989; Rikers et al. 2004) indicate that experts show higher accuracy rates and lower decisions times than novices. As one explanation of this finding it was argued that development of expertise is a (continuous) change from analytic to non-analytic cognitive processes (Eva 2004). Further, experts and novices seem to differ in the information gathering process to 'arrive at a diagnosis'. In particular, there is empirical evidence that experts used the same amount or less, but more relevant information as novices to make their decisions (Shanteau 1992). Barrows et al. (1982) also found that experts asked fewer questions and gathered less information than students, although they were more likely to provide

the correct diagnosis. However, none of the studies, which have examined the number of necessary diagnostic steps, considered decision times in their analyses.

Objectives of this study

The superiority of experts cannot always be demonstrated in traditional assessments that are used in competency-based medical education. For example, experienced doctors show significantly better performance on global scales than students or residents, while they scored significantly worse on the Objective Structure Clinical Examinations checklists (OSCEs, Boursicot et al. 2010; Ilgen et al. 2015), presumably because they took short cuts based on their experience (Hodges et al. 1999). One explanation for this finding is the use of a priori defined tasks in checklists. In typical performance-based medical examinations such as OSCEs, simulations, or workplace-based assessments mostly composite scores are created that integrate information about diagnosis accuracy, the number of diagnostic procedures taken, and the time needed for the problem solution. Thus, these aspects are difficult to separate in practice.

In this study we model these three aspects separately and adapt the approach suggested by Sherbino et al. (2012). We extend their research by empirically investigating the relationship between accuracy and decision speed in complex problems which more likely require System 2 than System 1 reasoning. The purpose of our study was to develop and validate a feasible and efficient computer-based testing instrument to assess clinical reasoning (ASCLIRE) by mimicking the entire clinical process closer than MCQs do. This process includes data acquisition, data interpretation, and data synthesis. In contrast to Sherbino et al. (2012), where the material presentation allowed participants to access all of the relevant data at once, in our test participants were asked to choose as many diagnostic measures as they deemed essential in the sequence of their choice. Thus, additionally to accuracy and decision time we consider the number of relevant gathered diagnostic information as a separate, third performance indicator of clinical reasoning.

Methods

Test development

Three board-certified academic anesthesiologists, two board-certified academic internists, and two educational psychologists developed six test cases. They reflected common causes for acute (three cases) and sub-acute (three cases) dyspnea based on real patients. The considered diagnoses were instable ventricular tachycardia, chronic obstructive pulmonary disease (COPD), pneumonia, intoxication, pulmonary artery embolism, and pulmonary edema.

All cases started with a written description of the respective clinical situation and a short video of the same male standardized patient in a hospital bed, displaying straightforward symptoms of the respective condition and make-up as required by the case.

In a second step, test participants could choose as many diagnostic measures as they deemed necessary in any sequence from a list of 36 diagnostic measures grouped into five categories (bedside tests, apperative diagnostics, patient history, previous reports, patient chart, and first measures; see Table 1). After clicking on the diagnostic measure, the result was presented to the participant either as text (e.g. pulse rate and strength), picture (e.g. ECG), or audio via headphones (e.g. heart auscultation). Most findings required

Table 1 Patient information and diagnostic measures available for each case

Bedside tests	Apperative diagnostics	Patient history	Previous reports	Patient chart	First measures
Take pulse	Chest X-ray	Current complaints	Previous ECG	Patient chart	Body position (4 to choose from)
Measure blood pressure	Arterial blood gas analysis	When did that start?	Previous chest X-ray		Apply oxygen
Measure breathing frequency	Chest ultrasound	Previous medical conditions?	Surgery protocol		Apply nitrospray s.l.
Auscultate lung	Transthoracic echocardiogram	Current medication?	Anesthesia protocol		Inhale beta-mimetics
Auscultate heart	Chest CT without contrast	Allergies?	Recovery room protocol		Establish venous access
Check pupils	Chest CT with contrast	Ever experienced that condition before?	Other previous medical reports		Call code blue
Measure blood glucose					
Oxygen saturation					
Take temperature					
3-Lead ECG					
12 Lead ECG					

CT computed tomography, *ECG* electrocardiogram

interpretation (e.g. heart sounds, ECG, chest X-ray). For radiology exams (e.g. ultrasound exams or CT scans) that appeared technically difficult to meaningful presentation in our test, the interpretation of a radiologist was given. An overview of considered diagnostic measures is provided in Table 1. Participants could give their final diagnosis at any time by clicking on the button "Diagnosis". In a third step they were asked to to provide their final diagnosis by selecting their diagnosis from a list of 20 diagnoses. The total time for problem solving was limited to 10 min per case. If the participant had a diagnosis in mind that was not in the list, he could choose the option "other". Furthermore, an alternative "I don't know" was offered.

Cases were presented in a random order, using the software Inquisit 2 (Millisecond Software 2010), which displays text, pictures, audios, and videos and also documents responses and reaction times. The presentation allows participants to access all of the relevant data in any order, thus allowing participants to rapidly access information they deem essential.

To assure the technical functionality of the test, we asked four medical students from the regular medical curriculum to think aloud while taking the test, focusing specifically on interface design and function. The think aloud protocols were screened by one researcher (WEH) for possible technical difficulties and opportunities for improvement.

Validation process

For test validation we employed a revised and extended framework of validity, proposed by Messick (1989) and introduced to the medical context by Downing (2003). This framework assumes that assessments are neither valid nor invalid and instead describes validity as a property of assessment scores. In this approach validity was conceptualized as a collection of evidence from a variety of different sources that support the proposed interpretations of test scores. In the present study we addressed the following aspects of validity: content evidence, internal structure, relations with other variables as well as evidence of consequences that are all briefly introduced in the next paragraph.

Content evidence can be evaluated by proving the representativeness of test and item specification; in our case the selection and construction of medical cases. The internal structure of assessments can be evaluated for example by providing information about item analysis or reliability scale analyses (e.g. internal consistencies). Furthermore, the empirical assessment structure can be examined using explanatory or confirmatory factor analysis. Relations with tests that assess a similar or identical construct inform about the convergent validity and are expected to be high, while correlations with assessments that assess different aspects reveal insights about divergent validity and are supposed to be lower than correlations reflecting convergent validity. Finally, evidence of consequences provides information about positive or negative consequences on the individual student level or on a more global level. Detailed information on the revised concept of validity can be found in Messick (1989) and American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999).

Student Participants

Participants were recruited from the students of the Reformed Medical Curriculum (RMC) at the Faculty of Medicine, Charité – Universitätsmedizin Berlin, Germany. The RMC was a 5-year parallel track curriculum that admitted 63 students per year from 1999 to 2009. Shortage of breath was taught in this curriculum during the first year (4 weeks

cardiovascular system, 1 week basic life support and 4 weeks of respiratory system), during the third year (1 week emergency medicine) and during the fourth year (4 weeks cardio-pulmonary diseases). Participation at the test was compulsory for all students taking the end-of semester examination of the RMC in year one to five. Informed consent was obtained from participants to match their ASCLIRE-results to their results in the end-of-semester examinations (OSCEs and MCQs) specific for each year of study). The Curriculum Committee and the Examination Board of the RMC (on 22.10.2009) as well as the Office of Data Protection at Charité (on 23.06.2009) gave their approval to perform the study. The Medical Ethical Review Committee also granted their approval on 07.09.2009 (EA1/170/09).

Since our test was developed as a normative assessment, we tested the cases against a reference panel using an expert sample of eight anesthesiologists and twelve internists with at least 3 years of professional experience since graduation. Seven experts were female. These experts were selected based on content expertise. The mean number of years practicing after graduation was 6.65 (SD = 2.93) and ranged from 3 to 12 years. The majority of experts had experience in working in the intensive care unit (19 of 20) and working as an emergency physician (15 of 20). Furthermore, 16 of 20 experts were involved in instructing emergency medicine or bedside teaching within the last two last years. Table 2 gives an overview of participants' demographics.

Procedure

ASCLIRE was administered in a computer laboratory over four consecutive days. Participants selected the testing date and time by signing up on a list. All students were offered a plenary presentation of the test's purpose and a live demonstration of a practice case 1 week prior to the study. A video of the plenary presentation was made available on the web to those who did not attend. Each test started with a 5 min video instruction on a computer, written instructions, and a practice case to familiarize the participants with the computer program. After viewing the video instruction, reading the written instructions, and completing the practice case, participants proceeded through the test materials, as described above and in Table 1.

After completing the test students provided feedback on the test. Students were asked whether they would consider the test an appropriate assessment in medical education

Table 2 Demographic characteristics of participants in the student and expert sample

Year	n	Age				Gender	
		Mean	Minimum	Maximum	SD	Female	Male
1	63	22.43	17	45	5.36	45	18
2	58	23.91	19	47	5.95	41	17
3	55	24.45	20	45	5.37	39	16
4	52	25.83	21	39	4.25	33	19
5	55	26.40	22	43	4.38	35	20
Total	283	24.52	17	47	5.29	193	90
Experts	20	–	–	–	–	7	13

n, number of participants in the (sub)sample; SD, standard deviation

practice and if they would wish for an equivalent test for other symptoms than shortage of breath.

Recorded Data

For each student demographical data (year of study, gender, age) as well as informed consent to match test results to MCQs and OSCE results were recorded. For each expert we documented years of practice since graduation, age, gender, discipline, and professional experience relevant to shortness of breath (emergency room, intensive care unit, pre-hospital emergencies).

Selected diagnostic measures, sequence of these measures, reaction times per measure, and selected diagnoses were recorded for each student and expert. Data were extracted from Inquisit's result files using R (R Development Core Team 2011) and transformed into conventional dataset format for further statistical analyses. The recorded reaction times were converted to response time in seconds.

Whereas the correct diagnosis was known a priori, the relevance of diagnostic measures—as a further performance indicator—was defined by the reference expert sample. A diagnostic measure was assumed to be relevant for the particular case, if at least 50 % of the expert sample have assessed that information.

Data analysis

The *diagnostic accuracy* was computed for each case separately using dichotomous answers (0/1)—indicating whether the correct diagnosis was provided. Other studies typically use 0/.5/1 for "wrong diagnoses, partly correct, correct" typically judged by experts. Our current approach is more conservative and appeared more appropriate to us, because the participants had to choose a diagnosis from a list of 20 possible diagnoses and did not enter it in a text field. Additionally, a percentage correct over all six cases was calculated. As a second indicator of student performance we considered the *number of relevant measures* that were requested for each case.

The temporal aspect of the diagnostic process was captured through the collection of reaction times. The total decision time per case (in seconds) from the first presentation of the patient video until the diagnosis selection may vary due to multiple factors (e.g. overall case difficulty or number of diagnostic measures taken in the case). Therefore, we instead have used the mean time per diagnostic measure per case as a "pure" indicator of decision time that captured how much time (on average) students needed for the integration of the results of any given diagnostic measure and thus informed whether new information was integrated fast or slow.

To evaluate the quality of the psychometric properties of the test we calculated item difficulties and item discrimination parameters in the total student sample for each of the six cases. Item difficulties were calculated as relative percentages of correct responses in the total student sample. Item discrimination parameters were computed as item-total-score-correlations as typically conceptualized in the classical test theory.

We used SPSS 21 for all statistical analyses except the structural equation modeling (SEM, see below). The general significance level for all analyses was defined at $p = .05$. For the group comparisons univariate ANOVAs were conducted. The performance of the experts was compared with performance of the total student group with regard to different outcomes (accuracy in the entire test, decision time, number of relevant measures) by calculating an univariate ANOVA with group membership (expert vs. student) as a

predictor variable. For the relations between different variables Pearson product-moment-correlations were calculated. All post hoc tests were computed using the Bonferroni correction at the significance level $\alpha = .05$. In addition, the effect size Cohen's d was reported for statistically significant comparisons between two groups. Cohen (1988) defined effect sizes about $d = .2$ as small; $d = .5$ as medium, and $d \geq .8$ as large effects. In order to investigate the relationship between accuracy and decision time we first calculated a Pearson-Product-correlation between mean accuracy (over all six cases) and mean decision time per diagnostic measure (over all six cases) for both students and experts.

In the second step we have modeled three distinct aspects of clinical reasoning as separate latent factors in a SEM. This method is appropriate to investigate the outlined research question for three reasons: First, this latent modeling approach allows simultaneously for flexible modeling of correlations between several latent variables representing different theoretical constructs. Second—in some way similar to G-studies (Bloch and Norman 2012)—this approach overcomes the problems of classical test theory and uses information from covariance matrices to account for the unreliability of single indicators by estimating the error variance that is not explained by the postulated latent factor (that is supposed to explain the observed measures). Third, in SEM it is possible to evaluate how well the empirical data fits the theoretical model using different model fit indices. For a general introduction to SEM see Raykov and Marcoulides (2000); for a brief introduction in the medical context please refer to Violato and Hecker (2007); more technical explanations can be found in Bollen (1989).

In our analysis we modeled the test structure empirically using the SEM approach by integrating the information on accuracy, mean decision time, and number of relevant measures simultaneously. Therefore, we defined one latent factor for the diagnostic accuracy with dichotomous answers (0/1) for specific cases as manifest indicators. For the latent second factor for decision time we did not use the total time required to solve the case, because as argued above, it comprises both information, namely amount of time and the number of requested measures. Instead we used the *mean decision time* per diagnostic measure, since we were interested in modeling distinct aspects of clinical reasoning that do not directly rely on the same information. For the third factor—representing the choice of relevant measures—the number of requested relevant measures (as defined by expert acquisition rate) was used as manifest indicators. Additionally, correlations between all three latent factors were postulated.

The SEM was estimated in Mplus 7.2 (Muthén and Muthén 1998–2010) using the weighted least standardized means and variance (WLSMV) estimator. WLSMV is a robust estimator which does not assume normally distributed variables, is based on tetrachoric correlations, and provides the best option for modeling categorical or ordered data (Brown 2006). As model fit indices, comparative fit index (CFI) and root mean square error of approximation (RMSEA) are reported, in addition to the χ^2 value. These measures of fit were included because the χ^2 value depends on sample size, where even small amounts of misfit can lead to significant χ^2 values when sample sizes are moderate to large (Chen 2007) and therefore can lead to misinterpretations (false model rejection or acceptance). For the RMSEA, values $\leq .05$ are taken to reflect a good fit, values between .05 and .08 an adequate fit. For CFI, values of .90 or higher are considered satisfactory fit, while values above .95 are considered excellent fit (Hu and Bentler 1999). The coefficient Ω for the latent reliability was calculated based on the estimated regression loadings in the SEM relying on tetrachoric correlations (McDonald 1999) and thus explicitly accounts for the unreliability of the single indicators. In particular, this coefficient is very effective for SEM models with dichotomous indicators, because the reliability of binary indicators usually

appears impaired when being estimated with traditional methods from classical test theory (e.g. Chronbach's alpha) due to their reduced variance.

Results

Number of relevant measures

Depending on the clinical case between 12 and 19 of 36 possible diagnostic measures were classified as relevant through their use by experts. The mean number of relevant measures across all six cases was 16.2.

Student performance

A total sample of 283 students participated in the study. In Table 3 the mean accuracy rates and the mean required solving time per case (in seconds) are displayed both for the total student sample and for each year separately. As a main pattern of diagnostic accuracy we observed a monotone improvement in diagnostic accuracy over years of study completed. ANOVA analysis revealed a significant increase in *diagnostic accuracy* over the course of the curriculum ($F(4,278) = 26.1$, $p < .001$). Moreover, post hoc tests between consecutive years revealed a significant increase between years with thematically relevant content—that is between years 1 (M = .37, SD = .21) and 2 (M = .48, SD = .24) with a medium effect of d = .49 and between years 3 (M = .57, SD = .21) and 4 (M = .69, SD = .21) with d = .57. Figure 1a shows the mean diagnostic accuracy over all six cases of students against year of study. Figure 1b shows the mean diagnostic accuracy for each case for medical students in different years and experts.

The differences in *total solution times* per case ($F(4,278) = .89$, $p = .47$) as well as in *mean decision time* per diagnostic measure ($F(4,278) = .32$, $p = .87$) were not statistically significant for different years of study.

We found significant differences in the *number of relevant measures* between different years in the student sample ($F(4,278) = 6.25$, $p < .05$). Post-hoc tests revealed significant

Table 3 Distribution of different performance indicators for students and experts

	n	Accuracy (% correct)			Mean reaction time per case (s)			Mean reaction time per diagnostic measure (s)			Mean number of relevant tests		
		M	SD	α	M	SD	α	M	SD	α	M	SD	α
Year 1	63	37.2	21.0	.22	184.4	47.9	.78	7.7	2.2	.87	11.7	2.1	.86
Year 2	58	48.6	24.1	.50	178.3	43.0	.74	7.9	1.7	.82	11.5	1.9	.83
Year 3	55	56.7	21.1	.24	188.9	46.5	.73	8.1	1.9	.83	12.2	1.9	.86
Year 4	52	69.3	20.7	.33	189.3	49.1	.75	7.9	2.2	.87	12.7	1.6	.74
Year 5	55	71.1	18.9	.16	194.0	49.5	.74	7.9	1.9	.83	13.0	1.7	.83
Students total	283	55.8	24.8	.48	186.8	47.2	.75	7.9	2.0	.84	12.2	1.9	.84
Experts	20	94.2	8.2	*	121.5	34.1	.71	6.00	1.4	.81	11.7	2.6	.86

n, number of participants in the (sub)sample; M, mean; SD, standard deviation; α, Chronbach's alpha

* Not meaningful due to the small size of the expert sample and very small variance in the mean accuracy scores

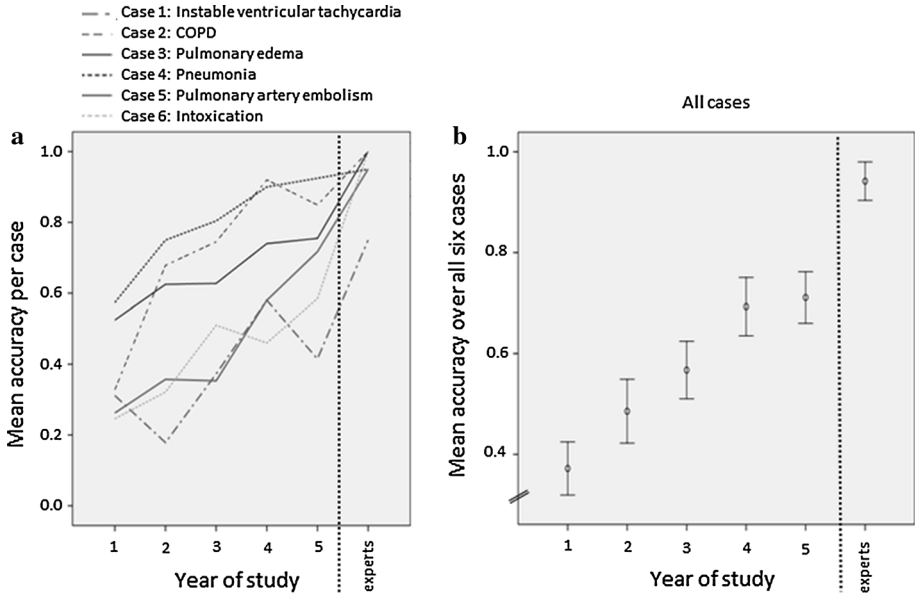

Fig. 1 **a** Mean accuracy per case for medical students in different years and experts. **b** Mean accuracy over all six cases for medical students (in different years) and experts

differences only between the years 1 and 5 ($d = .64$); years 2 and 4 ($d = .67$) as well as between years 2 and 5 ($d = .83$) (means and standard deviations are presented in Table 3. The Pearson-correlation between mean accuracy and mean decision time per diagnostic measure was not significant ($r = .03$, $p = .68$).

Expert performance and comparison with the student sample

Table 3 also presents mean *accuracy rates,* the number *of relevant diagnostic measures,* the mean *total solution time* per case, and the *mean time per diagnostic measure* for the expert sample. Experts show higher accuracy rates ($F(1,301) = 47.52$, $p < .001$, $d = 1.6$), need significantly less total time to finish the case ($F(1,301) = 36.93, p < .001, d = 1.4$) as well as less decision time per diagnostic measure ($F(1,301) = 18.07$, $p < .001$, $d = .98$) than students. Performance of experts and students did not differ with regard to the number of relevant measures ($F(1,301) = .987$, $p = .32$). The Pearson-correlation between mean accuracy and mean decision time per diagnostic measure was not significant ($r = -.22$, $p = .35$).

Psychometric properties of the test

Item difficulties and item discrimination parameters in the total student sample for each of the six cases are presented in Table 4. Mean item difficulties varied in the student sample between .36 and .78. The internal consistency (Cronbach's alpha) for the diagnostic accuracy over all six cases was .48. Item discrimination parameters ranged from .17 to .28 (with only one parameter lower than .20). Internal consistencies (Chronbach's alpha) were sufficient both for the mean time per diagnostic measure ($\alpha = .84$) and for the number of relevant measures ($\alpha = .84$).

Table 4 Item difficulty and item discrimination parameters for the diagnostic accuracy

Case	Item difficulty (relative frequencies of correct responses) Mean (SD)		Item discrimination (corrected item-test-correlation r_{it})
	Students (n = 283)	Experts (n = 20)	Students (n = 271)[a]
Case 1: instable ventricular tachycardia	.36 (.48)	.75 (.44)	.21
Case 2: COPD	.69 (.46)	1 (0)	.27
Case 3: pulmonary edema	.65 (.48)	1 (0)	.27
Case 4: pneumonia	.78 (.42)	.95 (.22)	.28
Case 5: pulmonary artery embolism	.45 (.50)	.95 (.22)	.26
Case 6: intoxication	.41 (.50)	1 (0)	.17

COPD chronic obstructive pulmonary disease

[a] For the calculation of corrected item-test-correlations 12 participants were excluded who did not work on all six cases

Testing the empirical test structure in SEM

For the estimated SEM all considered model fit indices indicated an excellent model fit ($\chi^2 = 152.2$, $df = 132$, $p = .11$; CFI = .983 was higher than .95; RMSEA = .023 was lower than .05). All regression loadings were significant and varied from .34 to .78. In general, the regression loadings for the accuracy factor were lower than the regression coefficients for the other two latent factors, presumably because the indicators for the accuracy factor were dichotomous. The correlation between the accuracy and decision time factor was not significant ($r = .04$, $p = .67$). Diagnostic accuracy was positively correlated with the number of relevant measures ($r = .51$, $p < .05$). Moreover, we obtained a small negative correlation between the number of relevant measures and the mean reaction time per diagnostic measure ($r = -.29$, $p < .05$). The SEM is illustrated in Fig. 2.

Latent reliability

Ω coefficient for the diagnostic accuracy based on tetrachoric correlations was with $\Omega = .63$ higher than Chronbach's $\alpha = .48$ reported above. In contrast, latent reliabilities for both the mean time per diagnostic measure ($\Omega = .85$) and for the number of relevant measures ($\Omega = .85$) were comparable to Chronbach's α.

Analysis of wrong diagnoses

The results show that wrong diagnoses provided by our sample were comparable to wrong diagnoses in practice. For example, the most common misdiagnosis in patients with instable ventricular tachycardia were myocardial infarction and atrial fibrillation with absolute tachyarrhythmia, which both are often associated with altered ECGs. The most common misdiagnosis for hypertensive pulmonary edema were again myocardial infarction (which in some cases causes pulmonary edema due to left ventricular failure) and pneumonia (which is often accompanied by lung edema).

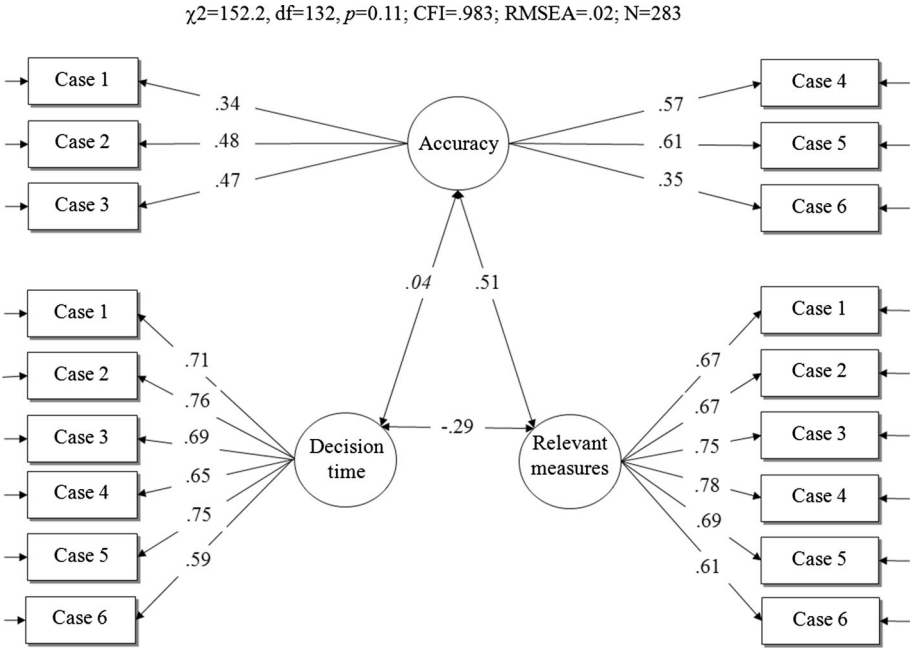

Fig. 2 Structural equation model for accuracy, mean decision time per diagnostic measure, and number of relevant measures in the student sample (n = 283). *Notes* df, degrees of freedom; CFI, comparative fit index; RMSEA, root mean square error of approximation; N, number of participants

Correlations with other assessment formats

We obtained informed consent from 250 students to match their ASCLIRE-results to their results in the end-of-semester examinations. Diagnostic accuracy correlates with results from the multiple-choice tests significantly only in year 2, 3, and 4 ($r = .37$; .41; .31; all $p < .05$) but not in year 1 and 5. Similarly, the results from the OSCE correlate with percentage of correct diagnosis only in year 5 ($r = .31$; $p = .34$). Also, MCQs and OCSE results in our data set correlated between $r = .15$ and $r = .57$ per year of study and were significant only for year 1 ($r = .57$) and year 3 ($r = .38$).

Acceptance of the test

72 % of the students agreed that such an assessment would be an appropriate assessment in medical education practice, whereas 15 % were not sure, and only 13 % did not agree with that statement. Further, 62 % of the students were interested in a test extension to other symptoms, whereas 28 % were not sure, and 10 % disagreed.

Discussion

The discussion is structured as follows: First, we provide a summary of the main results and critically discuss our findings concerning the reliability of the newly developed test. In the next part we discuss sources for the validity of test scores based on the results reported

above. In particular, we address content evidence, internal structure, relations with other variables as well as consequences evidence. In the last part we discuss the implication for practice and further research as well as limitations of the study and provide some conclusions.

Main results

As expected, the experts show higher accuracy rates and lower decision times than students, which are both typical findings to distinguish experts from novices (Bohle Carbonell et al. 2014; Ericsson et al. 2006; Lesgold et al. 1988). Another promising finding was the improvement of diagnostic accuracy with year of study similar to classic progress testing (Schuwirth and van der Vleuten 2012). Analysis of item difficulties and item discrimination parameters revealed that the newly developed instrument includes easy, moderate, and difficult cases. Using the structural equation modeling approach, we found an excellent fit for the model with three latent factors—diagnostic accuracy, decision time, and number of relevant diagnostic measures—with diagnostic accuracy showing a positive correlation with the number of relevant measures. In contrast, no significant correlation with decision time was found, demonstrating, that accuracy and decision time are two distinct aspects of clinical reasoning, at least when cases are sufficiently complicated to prompt system 2 reasoning. Furthermore, we found a small negative correlation between the decision time and the number of relevant measures, indicating that students who request more relevant measures needed on average less time to interpret the given diagnostic information.

Reliability

Reliability is nowadays typically conceptualized as one form of validity (Downing 2003) and is a necessary condition for the interpretation of validity evidence. When calculated using the traditional coefficient Chronbach's α, the reliability (i.e. internal consistency) for diagnostic accuracy was fairly low and needs further exploration.

According to the Spearman-Brown calculation of the number of cases required to reach a classic reliability of .7, 15 cases would be needed, which we consider as feasible. Such a test would take approximately two-and a half hours (or less, if the maximal time per case would be slightly reduced). Interestingly, Sherbino et al. (2012) report a comparably low internal consistency ($\alpha = .41$), although they had used 25 cases.

The internal consistency may be low due to the following two reasons. Chronbach's α is a function of the number of items, the mean inter-item-correlation (covariance), and item redundancy (Cortina 1993). Furthermore, it is possible that the internal consistency of a test is underestimated when the calculation is based on Pearson-product-moment correlations. The last assumption is supported by the result that the latent reliability (based on tetrachoric correlations) with $\Omega = .63$ is much higher than the reliability reported above. In general, pearson-product-moment-correlations lead to an underestimation of Cronbach's α, whereas tetrachoric correlations tend to overestimate it (Nunally 2007). Thus, the "true" reliability lies somewhere between the two reported values, more likely towards the upper bound estimate.

The reliabilities (Cronbach's α) calculated for each year separately appear even more problematic, which is probably due to the low number of students per group and constrained variance within these groups. This would be a serious shortcoming if the ASCLIRE test was used as a high-stake assessment for medical examinations by defining

separate norms for different years of study. However, in our research instrument we apply the performance of the experts as a norm and not the performance within the particular year of study. In addition as argued above, the internal consistencies per year will likely increase with greater sample sizes and larger amount of medical cases.

Sources for the validity of test scores

Content validity

The selection of our cases is comparable to those of other studies (Yudkowsky et al. 2009) of clinical reasoning, which implies that not only we but others have deemed similar cases representative of or otherwise relevant for the assessed content domain. Furthermore, the case construction was based on realistic medical cases and used real (anonymized) patient data (e.g. X-ray). In line with prior research (Lesgold et al. 1988) we found significant differences in the test performance between experts and students with large effect sizes, with experts solving diagnostic cases faster and more accurate than novices. The observed monotone increase in the test performance over the course of the curriculum with significant increases between the cohorts in the years with thematically relevant content is an even more promising result. As shown in Fig. 1a, b and Table 3 this linear trend can be demonstrated both for the overall test performance and for the specific cases. Moreover, acceptance of the new instrument by students was very high. While differences between novices and experts and participants' acceptance are commonly cited criteria for the validity of other testing formats, a monotone increase in performance over years of study is seldom reported. We argue that this finding is strong evidence for validity, indicating that the newly developed instrument might be appropriate to map student progress. Finally, another piece of evidence for content validity of ASCLIRE was gained through the analysis of wrong diagnoses that were comparable to wrong diagnoses in practice.

Internal structure

In terms of the internal structure we reported the psychometric properties of the test. In line with Messick's (1989) validity concept, we modeled the test structure empirically using the SEM approach by integrating the information on accuracy, reaction times, and number of relevant measures simultaneously.

We found evidence for three distinct aspects of clinical reasoning: diagnostic accuracy, decision time, and choice of relevant diagnostic information. The fact that we did not find a significant correlation between accuracy and decision time is very important, because we do not find evidence for the accuracy-speed trade off which would be indicated by a positive correlation between these two factors. Accuracy-speed trade off would imply that a lower time (hence greater speed) goes along with decreased accuracy. Our finding contradicts the results by Sherbino et al. (2012), who report a strong negative correlation between accuracy and response time. However, it is important to note that both studies are not directly comparable, since Sherbino et al. were mainly interested in studying system 1 reasoning. In contrast, we argue that ASCLIRE assesses system 2 reasoning, as suggested by the observation that students' and experts' response times in our test by far exceeded previously reported response times in presumably system 1 reasoning by Sherbino et al. (2012). Thus, our results indicate that—in particular when system 2 reasoning is involved—accuracy and decision time might reflect different aspects of clinical reasoning that are not related to each other.

Another interesting finding was a substantial positive correlation between diagnostic accuracy and the number of relevant measures. It indicates that a person who requests more relevant measures has a higher probability to get an accurate diagnosis. This finding aligns with prior research demonstrating that experts tend to gather less but more relevant information than novices (Shanteau 1992) and might mean that the choice of particular measures appears not random but requires intentional control. This additionally supports the assumption that it is very likely that the involved cognitive processes are analytic (system 2) rather than non-analytic (system 1)—at least in the student sample. This thesis is also supported by the small increase of the number of relevant diagnostic measures over the years of study. Given the substantially lower total decision times per case in the expert sample, a more detailed analysis of solution strategies in the expert sample is required to investigate to which extent the diagnostic process requires conscious control for experts.

In a nutshell, these results imply that high diagnostic accuracy—as one important criterion of any medical decision—is potentially influenced by two factors: time required for the decision process and the amount of acquired relevant diagnostic information. The number of relevant measures was associated with higher diagnostic accuracy, whereas—in contrast to previous studies—diagnostic accuracy and decision time were not related.

Relations with other variables

To test the construct validity, we calculated correlations with variables outside the test under study, namely MCQs and OCSE results and did not find strong correlations. In fact, ASCLIRE correlated about as strong to MCQs and OSCE results as those two correlate among each other. As most would agree that MCQs and OSCE measure different aspects of clinical reasoning (declarative and procedural knowledge respectively), we would argue that ASCLIRE measures a third construct. These results should be interpreted with caution, because the ASCLIRE test included medical cases around shortage of breath only, while MCQs and OCSEs only partly assessed this domain (if they assessed it at all) and covered different content domains. Therefore, we assume that the correlations and their magnitude reflect at least partly relevant aspects of general intelligence, meaning that people who perform well in one (medical) assessment will likely perform well in another, regardless of content domain. Correlation studies investigating the relationship between e.g. scores from a script concordance test, as one established measure of clinical reasoning, and ASCLIRE could provide evidence for convergent and divergent validity of scores in the presented test.

Consequences evidence

The last source of validity derives from students' consequences from taking the test. About 10 % of students, taking the test, asked for and attended a seminar discussing different diagnostic approaches towards patients with shortage of breath. Based on the individual feedback students received from taking the test, we in-depth discussed the available diagnostic measures, their importance and information content along with potential diagnostic considerations. Thus, the participation in the test seems to evoke interest at least for a small sub-sample of the students.

Implication for practice and further research

As accuracy and decision time seem to be distinct aspects of clinical reasoning, it appears desirable to prepare and train medical students with respect to both aspects of clinical reasoning in medical education. In contrast to model-based tests, ASCLIRE does not make assumptions about the underlying cognitive processes a priori. Nevertheless as argued above, it is very likely that explicitly asking students to thoroughly assess the virtual patients and take all diagnostic measures they deem relevant evokes an analytic processing (using system 2 reasoning) and that only few, if any students, employed faster pattern recognition or other system 1 approaches. To shed light on the involved cognitive processes, conducting think–aloud studies might be promising.

Validity, as it is conceptualized throughout this study, is not provable but much rather an informed consideration of arguments in favor and in opposition of it. Moreover, validity is contextually bound, i.e. an instrument considered valid in one context (e.g. undergraduates) is not necessarily valid in another (e.g. postgraduates) (Downing and Haladyna 2009). Further studies of the presented instrument in different contexts are required to conclude on its transferability.

Although the presented test has some promising psychometric properties, the current findings are limited to the domain "shortness of breath". Since clinical reasoning at least partly depends on the clinical context (Durning et al. 2012; Eva 2003; Wimmers et al. 2007), an extension of the test content to other medical domains or symptoms is desirable. It would also help to gather further evidence for convergent and divergent validity of the test scores.

Limitations

Our study has several limitations. Foremost, it is a cross sectional study that does not investigate the individual development of clinical reasoning. Nevertheless, preliminary results indicate a significant improvement of clinical reasoning during the undergraduate medical education, which is a promising and not trivial result.

We are currently not able to present strong arguments for evidence of consequences, which is partly due to the pilot character of the present study and the difficulty in specifying such consequences. A recent review of validity arguments for simulation based assessment has found very few sources of this validity evidence being reported (Cook et al. 2014).

Furthermore, participation was compulsory for all students, which on the one hand may lead to "click-trough" participation, which could in turn water the results down. We controlled the data to determine such pattern (e.g. by examining cases with extremely short reaction times). On the other hand we argue that the compulsory participation is rather a strength of the study, because we avoid selection bias in the sample.

In addition, the diagnostic measures and the possible diagnosis were predetermined and not implemented as free text fields. This was done to improve the technical handling of the test. For a more detailed discussion of the implications of different question formats for eliciting thinking processes see Schuwirth et al. (2001).

Moreover, it is important to note that the definition of relevant measures is to a certain extent arbitrary. We have decided to choose a rather conservative criterion, which was that the majority of experts had chosen that measure. Another possibility would have been to weight the choice of each measure by the percentage of experts that have used that diagnostic information (as commonly done in script concordance testing).

Finally, it is possible, that the correlations between MCQs, OSCEs, and ASCLIRE were underestimated, because these tests are not necessarily comparable with regard to the content assessed in the particular tests.

Conclusion

ASCLIRE is a norm-referenced computer-based instrument to assess clinical reasoning. In addition to comparable existing tests that integrate the entire diagnostic process, we consider decision time as another important performance indicator beneath diagnostic accuracy and the number of relevant diagnostic measures. The presented instrument meets to the greatest extent traditional psychometric requirements in terms of reliability and validity. Using a structural equation modeling approach, we found evidence for three partly related but still distinct aspects of clinical reasoning: diagnostic accuracy, decision time, and choice of relevant diagnostic information. First results from this cross-sectional study suggest that this test instrument is suitable to describe development of clinical reasoning during the years of undergraduate medical education.

Acknowledgments The authors would like to acknowledge Tim Ullrich for his support in data acquisition, Clemens De Grahl († 20.03.2012) and Eva Kornemann for their contribution to the development of test cases, Raimund Senf for his support in acquiring expert test data, Torsten Schröder for his contribution to the initiation of the study, and Tamara Pace Ross for proofreading.

Conflict of interest The authors report no conflicts of interest.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA, APA, NCME.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and Investigative Medicine, 5*(1), 49–55.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5, Supplement), S2–S23. doi:10.1016/j.amjmed.2008.01.001.

Bloch, R., & Norman, G. R. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher, 34*(11), 960–992. doi:10.3109/0142159X.2012.703791.

Bohle Carbonell, K., Stalmeijer, R. E., Könings, K. D., Segers, M., & Van Merriënboer, J. J. G. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review,*. doi:10.1016/j.edurev.2014.03.001.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boursicot, K. A., Roberts, T. E., & Burdick, W. P. (2010). Structured assessments of clinical competence. In T. Swanwick (Ed.), *Understanding medical education: Evidence, theory, and practice* (1st ed., pp. 246–258). New Jersey: Wiley-Blackwell.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guildford.

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., et al. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education, 46*(5), 454–463. doi:10.1111/j.1365-2923.2012.04242.x.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education, 19*, 233–250. doi:10.1007/s10459-013-9458-4.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Croskerry, P. (2009). Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. *Advances in Health Sciences Education, 1*, 27–35.

Croskerry, P., & Norman, G. R. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121*, 24–29.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837. doi:10.1046/j.1365-2923.2003.01594.x.

Downing, S. M., & Haladyna, T. M. (2009). Validity and its threats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 21–56). New York: Routledge.

Durning, S., Artino, A., Boulet, J., Dorrance, K., van der Vleuten, C., & Schuwirth, L. (2012). The impact of selected contextual factors on experts' clinical reasoning performance (Does context impact clinical reasoning performance in experts?). *Advances in Health Sciences Education, 17*(1), 65–79. doi:10.1007/s10459-011-9294-3.

Elstein, A. S. (2009). Thinking about diagnostic thinking: A 30-year perspective. *Advances in Health Sciences Education, 14*(1), 7–18. doi:10.1007/s10459-009-9184-0.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *JAMA: Journal of the American Medical Association, 287*(2), 226–235. doi:10.1001/jama.287.2.226.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.

Eva, K. W. (2003). On the generality of specificity. *Medical Education, 37*(7), 587–588. doi:10.1046/j.1365-2923.2003.01563.x.

Eva, K. W. (2004). What every teacher needs to know about clinical reasoning. *Medical Education, 39*(1), 98–106. doi:10.1111/j.1365-2929.2004.01972.x.

Gandhi, T. K., Kachalia, A., Thomas, E. J., Puopolo, A. L., Yoon, C., Brennan, T. A., & Studdert, D. M. (2006). Missed and delayed diagnoses in the ambulatory setting: A study of closed malpractice claims. *Annals of Internal Medicine, 45*, 488–496.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493–1499. doi:10.1001/archinte.165.13.1493.

Hays, R. (2014). The potential impact of the revision of the Basic World Federation Medical Education Standards. *Medical Teacher, 36*, 459–462.

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

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*, 1–55.

Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education, 49*(2), 161–173. doi:10.1111/medu.12621.

Kachalia, A., Gandhi, T. K., Puopolo, A. L., Yoon, C., Thomas, E. J., Griffey, R., et al. (2007). Missed and delayed diagnoses in the emergency department: A study of closed malpractice claims from 4 liability insurers. *Annals of Emergency Medicine, 49*(2), 196–205. doi:10.1016/j.annemergmed.2006.06.035.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342).

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan Publishing Co, Inc; American Council on Education.

Millisecond Software. (2010). *Inquisit* (*Version 2*). Seattle: Washington.

Muthén, L. K., Muthén, B. (1998–2010). *Mplus* (*Version 7.2*). Los Angeles, CA: Muthén & Muthén.

Newman-Toker, D. E., & Pronovost, P. J. (2009). Diagnostic errors: The next frontier for patient safety. *JAMA: Journal of the American Medical Association, 301*(10), 1060–1062. doi:10.1001/jama.2009.249.

Norman, G. R. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418–427. doi:10.1111/j.1365-2929.2005.02127.x.

Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education, 2*(2), 173–184.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology, 125*(8), 1063–1068.

Norman, G. R., Sherbino, J., Dore, K., Wood, T., Young, M., Gaissmaier, W., et al. (2014). The etiology of diagnostic errors: A controlled trial of system 1 versus system 2 reasoning. *Academic Medicine, 89*(2), 277–284.

Nunally, J. C. (2007). *Introduction to psychological measurement*. New York: McGraw-Hill.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rikers, R. M., Loyens, S. M., & Schmidt, H. G. (2004). The role of encapsulated knowledge in clinical case representations of medical students and family doctors. *Medical Education, 38*(10), 1035–1043.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). The use of progress testing. *Perspectives on Medical Education, 1*(1), 24–30. doi:10.1007/s40037-012-0007-2.

Schuwirth, L. W. T., Verheggen, M. M., Van Der Vleuten, C. P. M., Boshuizen, H. P. A., & Dinant, G. J. (2001). Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education, 35*(4), 348–356. doi:10.1046/j.1365-2923.2001.00771.x.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica, 81*(1), 75–86. doi:10.1016/0001-6918(92)90012-3.

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., & Norman, G. R. (2012). The relationship between response time and diagnostic accuracy. *Academic Medicine, 87*(6), 785–791. doi:10.1097/ACM.0b013e318253acbd.

Singh, H., Thomas, E. J., Petersen, L. A., & Studdert, D. M. (2007). Medical errors involving trainees: A study of closed malpractice claims from 5 insurers. *Archives of Internal Medicine, 167*(19), 2030–2036. doi:10.1001/archinte.167.19.2030.

Swing, S. (2007). The ACGME outcome project: Retrospective and prospective. *Medical Teacher, 29*, 648–654.

Violato, C., & Hecker, K. G. (2007). How to use structural equation modeling in medical education research: A brief guide. *Teaching and Learning in Medicine, 19*(4), 362–371.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67–85. doi:10.1016/0001-6918(77)90012-9.

Wimmers, P., Splinter, T. W., Hancock, G., & Schmidt, H. (2007). Clinical competence: General ability or case-specific? *Advances in Health Sciences Education, 12*(3), 299–314. doi:10.1007/s10459-006-9002-x.

Yudkowsky, R., Otaki, J., Lowenstein, T., Riddle, J., Nishigori, H., & Bordage, G. (2009). A hypothesis-driven physical examination learning and assessment procedure for medical students: Initial validity evidence. *Medical Education, 43*(8), 729–740. doi:10.1111/j.1365-2923.2009.03379.x.