



Measuring Faculty Reflection on Adverse Patient Events: Development and Initial Validation of a Case-Based Learning System

Christopher M. Wittich, MD, PharmD¹, Francisco Lopez-Jimenez, MD², Lindsay K. Decker, BS³, Jason H. Szostek, MD¹, Jayawant N. Mandrekar, PhD⁴, Timothy I. Morgenthaler, MD⁵, and Thomas J. Beckman, MD¹

¹Department of Internal Medicine, Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA; ²Department of Internal Medicine, Division of Cardiovascular Diseases, Mayo Clinic College of Medicine, Rochester, MN, USA; ³Department of Internal Medicine, Information Technology, Mayo Clinic College of Medicine, Rochester, MN, USA; ⁴Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA; ⁵Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA.

BACKGROUND: Critical reflection by faculty physicians on adverse patient events is important for changing physician's behaviors. However, there is little research regarding physician reflection on quality improvement (QI).

OBJECTIVE: To develop and validate a computerized case-based learning system (CBLS) to measure faculty physicians' reflections on adverse patient events.

DESIGN: Prospective validation study.

PARTICIPANTS: Staff physicians in the Department of Medicine at Mayo Clinic Rochester.

MAIN MEASURES: The CBLS was developed by Mayo Clinic information technology, medical education, and QI specialists. The reflection questionnaire, adapted from a previously validated instrument, contained eight items structured on five-point scales. Three cases, representing actual adverse events, were developed based on the most common error types: systems, medication, and diagnostic. In 2009, all Mayo Clinic hospital medicine, non-interventional cardiology, and pulmonary faculty were invited to participate. Faculty reviewed each case, determined the next management step, rated case generalizability and relevance, and completed the reflection questionnaire. Factor analysis and internal consistency reliability were calculated. Associations between reflection scores and characteristics of faculty and patient cases were determined.

KEY RESULTS: Forty-four faculty completed 107 case reflections. The CBLS was rated as average to excellent in 95 of 104 (91.3%) completed satisfaction surveys. Factor analysis revealed two levels of reflection: *Minimal* and *High*. Internal consistency reliability was very good (overall Cronbach's $\alpha=0.77$). Item mean scores ranged from 2.89 to 3.73 on a five-point scale. The overall reflection score was 3.41 (standard deviation 0.64). Reflection scores were positively associated with case generalizability ($p=0.001$), and case relevance ($p=0.02$).

Electronic supplementary material The online version of this article (doi:10.1007/s11606-010-1548-x) contains supplementary material, which is available to authorized users.

Received June 11, 2010

Revised September 29, 2010

Accepted October 7, 2010

Published online October 27, 2010

CONCLUSIONS: The CBLS is a valid method for stratifying faculty physicians' levels of reflection on adverse patient events. Reflection scores are associated with case generalizability and relevance, indicating that reflection improves with pertinent patient encounters. We anticipate that this instrument will be useful in future research on QI among low versus high-reflecting physicians.

KEY WORDS: assessment; medical education; quality improvement; reflection; validation study; adverse events.

J Gen Intern Med 26(3):293-8

DOI: 10.1007/s11606-010-1548-x

© Society of General Internal Medicine 2010

BACKGROUND

The Institute of Medicine's landmark report, *To Err is Human*, focused attention on patient safety.¹ To improve patient care, many healthcare organizations review sentinel events,² study patient mortality data,³ and convene morbidity and mortality conferences.⁴ The goal of these endeavors is to prevent adverse events by identifying root causes and solutions. However, there is little conclusive research regarding solutions that influence physicians' practice behaviors.

Electronic case-based curricula on patient safety and systems-based practice have been studied among residents and faculty members including emergency medicine, family medicine, and internal medicine.⁵⁻⁸ All of these studies have shown that case-based curricula improve physician knowledge. However, we are unaware of research about *reflection on adverse events* to change future practice behaviors.

Kirkpatrick's outcomes hierarchy is a common framework for education intervention studies.^{9,10} The following are the hierarchical levels: 1) Reaction (learner satisfaction), 2) Learning (acquired knowledge or skills), 3) Behavior (transferring learning to the workplace), and 4) Results (patient outcomes).^{9,10} When progressing up this hierarchy from Reaction to Results, the outcomes become increasingly meaningful, yet more difficult to measure with respect to feasibility and methodological rigor. Although most education interventions

involve outcomes at the levels of Reaction and Learning,¹¹ experts have observed that Behavior outcomes strike the optimal balance between feasibility and meaningfulness.¹² One way to change physicians' behaviors is to encourage critical reflection on adverse patient events.¹³ Critical reflection is defined as "a meta-cognitive process that occurs before, during and after situations with the purpose of developing greater understanding of both the self and the situation, so that future encounters with the situation are informed from the previous encounters."¹⁴ Remarkably, validated systems to measure physician reflection on adverse events have not been described.

Therefore, we created an electronic system for presenting cases of adverse patient events in order to stimulate and measure faculty physician reflection on those events. In order to stratify physicians into low versus high levels of reflection, we created a measure of reflection based on a previously validated instrument, which separates learners into four increasing levels of reflection ranging from habitual action (no reflection) to critical reflection (highest reflection). We hypothesized that the degree of physician reflection would be associated with characteristics of physicians (e.g., age, gender) and patient cases (e.g., event relevance). The objectives of this study were to 1) develop an electronic case-based learning curriculum regarding actual systems failures and adverse patient events, 2) create a reflection instrument based on previous research^{15,16} and validate it in a population of faculty physicians, and 3) determine associations between physician reflection and characteristics of physicians and patient cases.

METHODS

Study Population

This study was conducted in 2009 and involved department of internal medicine faculty at the Mayo Clinic Rochester. All faculty members in the divisions of hospital internal medicine (generalists), pulmonary diseases and critical care medicine (specialists), and non-interventional cardiovascular diseases (sub-specialists) were invited to participate. This study was deemed exempt by the Mayo Institutional Review Board.

Case-Based Learning System (CBLS)

The CBLS (available in an online Appendix) was developed by Mayo Clinic specialists in the sections of information technology, medical education, and quality improvement (QI). The system was modified based on multiple rounds of feedback from small test groups. In August, 2009, the final system was administered to the study sample via e-mail messages that contained links to the web-based system, which in turn presented three patient cases. Participants who had not responded were sent two e-mail reminders. Participants were instructed to read each case and answer the corresponding multiple choice question following each case. After completing all three cases, the participants were asked to complete the reflection questions. Data collection was completed within four weeks.

After the participants accessed links to the CBLS, they reviewed the patient cases and answered multiple choice questions (MCQ) that tested understanding of the healthcare system (not understanding of medical knowledge), and required commitment to the next step in management. Subsequently, participants were provided with the case resolution where the adverse event was revealed. Participants then rated their perception of the case's generalizability and relevance, and completed the reflection questionnaire. Finally, the participants read an informational case discussion and completed a satisfaction survey, which asked whether the CBLS met their needs for CME structured on a five-point scale (unsatisfactory, needs improvement, average, above average, excellent).

Three cases, representing real adverse events at the Mayo Clinic, were selected based on the most common error types (systems, medication, and diagnostic) encountered by internal medicine physicians.¹⁷ The systems error case involved a hospitalized patient who developed a ST-elevation myocardial infarction. The team did not follow hospital protocol to activate the cardiac catheterization laboratory, which led to delayed treatment and a poor patient outcome. The medication error case involved a patient whose home medications were not reconciled at the time of admission to the hospital. This error resulted in a potassium overdose, intensive care unit transfer, and prolonged hospital stay. The failure to diagnose case involved an elderly woman who had provided a urine culture on the day of hospital discharge. After discharge the culture turned positive, but no one reviewed the result. The patient returned to the hospital three days later with urosepsis. All cases were reviewed and edited by a generalist (author CMW), a specialist (author FLJ), and a systems expert (author TIM).

The reflection instrument used in the CBLS was adapted from a previously validated tool by Kember *et al.*, which comprised four levels of reflection: habitual action, understanding, reflection, critical reflection.^{15,16} *Habitual action* is a perfunctory feat that through repetition has become automatic.^{15,16} *Understanding* is using existing knowledge without critically appraising that knowledge.^{15,16} *Reflection* is exploration of past experiences to develop new understandings.^{15,16,18} *Critical reflection* is a deeper form of reflection where a person's perspective is changed.^{15,16} We adapted Kember's tool to our setting by creating eight items (two for each level of reflection) structured on five-point Likert scales (1=Disagree, 2=Disagree with reservation, 3=Neutral, 4=Agree with reservation, 5=Agree). (See Table 1).

Data Analysis

Satisfaction survey, multiple choice, and item score data were presented using standard descriptive statistics. Confirmatory factor analysis with Varimax rotation was performed to determine dimensions of physician reflection on adverse events. Specifically, we wished to confirm whether items clustered into conceptual groups representing low versus high reflection, as previously demonstrated by Kember.^{15,16} Factors were extracted using the minimal proportion criteria, which is the proportion of the common variance (defined by the sum of communality estimates) that is explained by successive factors. In this study, we established the threshold at 90% of the common variance, and factors were extracted until the sum of

Table 1. Faculty Physician Reflection on Adverse Patient Events: Item Loadings

Item	Reflection Level ^{15,16}	Item loadings	
		Factor 1	Factor 2
1. When I do activities like in this case, I complete them without thinking about what I am doing	Habitual Action	0.085	0.528
2. If I follow what the case says, I do not have to think too much	Habitual Action	-0.040	0.560
3. This case required me to understand the concepts taught by the module	Understanding	0.792	0.106
4. I need to understand the material taught by the case in order to perform practical tasks	Understanding	0.605	0.141
5. This case caused me to reflect on my actions to see whether I could improve on what I do	Reflection	0.864	0.049
6. This case made me think over what I have been doing and consider alternative ways of doing it	Reflection	0.758	0.057
7. As a result of this case, I will change my normal way of doing things	Critical Reflection	0.628	-0.120
8. During this case, I discovered faults in what I had previously believed to be right	Critical Reflection	0.570	-0.025

Column 2 shows each of the original Kember levels of reflection, and reveals how these correspond to the levels of minimal (Factor 2) and high (Factor 1) reflection found in the current study

Eigenvalues for the retained factors exceeded 90% of the common variance, defined as the sum of the initial communality estimates. The final model was confirmed by examining the scree plot, which shows relative magnitudes of the factors' Eigenvalues and can aid in determining how many factors to retain by inspecting to see where the declining Eigenvalues level off.¹⁹ Items with factor loadings ≥ 0.30 were retained. Internal consistency reliability for items comprising each factor and overall were calculated using Cronbach's coefficient α , where $\alpha > 0.7$ is considered acceptable.^{19,20}

Associations between reflection scores and learner or case variables were determined. Overall reflection scores were reported as the mean and standard deviation of all eight instrument items. For reporting and all associations calculations, values for items comprising Factor 2 were reverse-scored, given the negative phrasing of these items. Case relevance (yes/no), event generalizability (yes/no), event preventability (yes/no), event root cause (personal/system), physician gender (female/male), and multiple choice answer correct (yes/no) were treated as binary variables. Event severity (near miss, minor impact, moderately severe impact, severe impact, death) was treated as an ordinal scale. Physician age was treated as a continuous variable. ANOVA and linear regression analysis were used to determine associations between overall reflection score and categorical or continuous variables, respectively. Statistical significance was set at $p < 0.05$. Statistical

analyses were performed using SAS version 9.2 (SAS Institute Inc., Cary, NC).

RESULTS

Faculty Satisfaction and Responses to CBLS Cases

The CBLS was accessed by 44 (38%) of the 116 faculty physicians invited to participate. The 44 participants were given three cases apiece and completed a collective total of 107 case reflections. The CBLS was rated as average to excellent in 95 of 104 (91.3%) completed satisfaction surveys. In the ST-segment elevation myocardial infarction case, which asked how to activate the cardiac catheterization laboratory, 7 of 29 (24.1%) identified the correct answer. In the potassium overdose case, which asked about the current policy regarding medication reconciliation, 15 of 38 (39.5%) identified the correct answer. In the missed urine culture case, which asked about the protocol regarding who is responsible for pending laboratory data at the time of patient discharge, 25 of 40 (62.5%) identified the correct answer.

Reflection Instrument Validation

Factor analysis revealed a two-dimensional model for measuring faculty physicians' reflections on adverse patient events. The identified factors were: 1) *Minimal Reflection* (items 1 and 2), and *High Reflection* (items 3 through 8). These factors support Kember's model by distinguishing between low and high levels of reflection (Table 1). Specifically, the *Minimal Reflection* factor was comprised of two items corresponding to Kember's Habitual Action level, and the *High Reflection* factor was comprised of items that correspond to Kember's Understanding, Reflection, and Critical Reflection levels. Overall, the extracted factors accounted for 100% of the shared variance among the original variables.

Item mean scores ranged from 2.89 to 3.73 on a five-point scale. The overall reflection score was 3.41 (standard deviation 0.64). Remarkably, mean items scores were highest for items representing the lowest levels of reflection (e.g., "When I do activities like in this case, I complete them without thinking about what I am doing;" mean score=3.66), and lowest for items representing the highest levels of reflection (e.g., "During this case, I discovered faults in what I had previously believed to be right;" mean score=2.89), indicating that, on average, high level reflection was less commonly achieved in this study sample (Table 2). Regarding internal consistency reliability, Cronbach's α for Factor 1 was 0.85, for Factor 2 was 0.58, and overall was 0.77 (Table 2).

Reflection Score Associations

ANOVA analysis indicated that reflection scores were associated with physicians' perceptions of case relevance ($p=0.02$) and event generalizability ($p=0.001$). There were no statistically significant associations between physicians' reflection scores and event severity, event preventability, root cause, physician

Table 2. Faculty Physician Reflection on Adverse Patient Events: Factors, Mean Scores and Reliability

Items	Mean*	SD
Factor 2 – Minimal Reflection (Cronbach's $\alpha=0.58$)		
When I do activities like in this case, I complete them without thinking about what I am doing	3.66	1.12
If I follow what the case says, I do not have to think too much	3.46	1.09
Factor 1 – High Reflection (Cronbach's $\alpha=0.85$)		
This case required me to understand the concepts taught by the module	3.73	1.13
I need to understand the material taught by the case in order to perform practical tasks	3.47	0.97
This case caused me to reflect on my actions to see whether I could improve on what I do	3.71	0.98
This case made me think over what I have been doing and consider alternative ways of doing it	3.47	1.13
As a result of this case, I will change my normal way of doing things	2.93	1.03
During this case, I discovered faults in what I had previously believed to be right	2.89	1.18
Overall – All 8 items (Cronbach's $\alpha=0.77$)	3.41	0.64

*Items comprising Factor 2 were reverse score so that all item values were positive

SD = standard deviation; Cronbach's α represents internal consistency reliability for the items comprising each factor and overall

gender, physician age, or multiple choice answer response (Table 3).

DISCUSSION

To our knowledge, this is the first study of a case-based learning system for measuring faculty physicians' reflections on adverse patient events. The CBLS reflection instrument scores were reliable and stratified faculty members across two levels of reflection, which should prove useful for future research regarding QI among low versus high-reflecting physicians. Furthermore, reflection scores were positively associated with case generalizability and case relevance, indicating that reflection improves with the use of cases that represent actual patient encounters.

We found that reflection scores correlated positively with case generalizability and relevance. These findings suggest that cases stimulating the richest reflections are those having obvious bearing on one's own practice. In this study, generalizability was enhanced by selecting cases which have been shown to occur commonly in the setting of adverse events.¹⁷ Relevance was optimized by selecting cases that represented real adverse events at the Mayo Clinic. Physicians may be most comfortable when contemplating familiar ideas, even though, arguably, healthcare improvement is driven by reflecting on the unknown. While cases in this study were comprised of familiar content, physicians generally scored poorly regarding knowledge of critical management steps. Therefore, we intend to use the CBLS to electronically disseminate adverse patient events to the entire faculty, with the aim of educating faculty members regarding key quality initiatives.

Reflection has been described as a process of thinking critically about all aspects of a situation, including the self, and has even been described as "thinking about thinking."¹⁴ Schon further observes that artful practitioners representing

Table 3. Associations between Reflection Scores and Characteristics of Faculty and Patient Cases

Variable	Percentage	Mean Reflection Score*	SD	p value**
Physician Perception of Case Relevance				0.02
Yes	84.1%	3.47	0.62	
No	15.9%	3.10	0.68	
Physician Perception of Event Severity				0.39
No Impact to Patient (Near Miss)	2.8%	2.79	0.71	
Minor Impact to Patient	1.9%	2.94	0.44	
Moderate Impact to Patient	68.2%	3.47	0.57	
Severe Impact to Patient	27.1%	3.37	0.78	
Patient Death	0%			
Physician Perception of Case Generalizability				0.001
Likely or Moderately Likely	86.0%	3.49	0.60	
Not Likely	14.0%	2.93	0.68	
Physician Perception of Event Preventability				0.52
Preventable	99.1%	3.42	0.64	
Not Preventable	0.9%	3.00	na	
Physician Perception of Root Cause				0.75
Personal/Team Factor	53.3%	3.39	0.70	
System Factor	46.7%	3.44	0.57	
Physician Gender				0.63
Female	10.3%	3.38	0.66	
Male	86.9%	3.47	0.59	
Gender missing	2.8%			
Physician Age				0.33
30–39 years	12.1%	3.38	0.60	
40–49 years	25.2%	3.33	0.54	
50–59 years	43.9%	3.64	0.60	
≥60 years	6.5%	3.18	0.58	
Age missing	12.1%			
Multiple Choice				0.61
Question Answer				
Correct	43.9%	3.39	0.66	
Incorrect	56.1%	3.45	0.61	

*Items comprising Factor 2 were reverse scored so that all item values were positive

**Association between reflection score and the variable using ANOVA or linear regression analysis, where appropriate

diverse disciplines share the ability to deal with unique circumstances by having "reflective conversations with the situation."²¹ Previous studies on reflection among healthcare workers exist. The Mayo Evaluation of Reflection on Improvement Tool (MERIT) assesses resident physicians' reflections on adverse events encountered in practice.²² However, MERIT reflections are scored by external raters, and are thus cumbersome and biased. The Kember instrument, which was the basis for this study's assessment tool, measures health sciences students' reflections on their courses.^{15,16} The Groningen Reflection Ability Scale (GRAS)²³ and the Self-Reflection and Insight Scale (SRIS)²⁴ are additional measures of reflection in educational settings. The Mayo CBLS reflection assessment instrument adds to

this body of literature, because it specifically measures faculty physicians' reflections on adverse patient events. Additionally, we found that faculty members scored lowest on critical reflection compared with the other categories. This finding provides preliminary evidence that, among faculty physicians, the highest level of reflection on QI may be difficult to achieve.

Our measure of faculty physician reflection on adverse events is supported by validity evidence. An established validity paradigm states that construct validity is upheld by evidence from the following sources: content, response process, internal structure, relations to other variables (criterion), and consequences.^{25–30} In this study, *content evidence* draws from items that were created based on a previously validated reflection assessment instrument,^{15,16} findings from our prior research,²² and revision by a panel of experts with experience in teaching and measuring reflection and QI. *Internal Structure* evidence is supported by factor analysis showing a two-dimensional assessment of physician reflection that generally verifies the reflection levels found in the original Kember instrument,^{15,16} and by excellent overall internal consistency reliability. *Relations to other variables* (criterion) evidence is established by associations between reflection scores and other meaningful variables including case characteristics. Our previous review of the literature indicated that the aforementioned sources of validity evidence are the categories most commonly reported in the literature.²⁹ Nonetheless, in the future it will be important to establish *Consequences* evidence by determining whether our assessment of reflection actually has an impact on faculty physician's abilities to improve patient care through effective QI endeavors.

This study has limitations. It was conducted at a single academic institution, which may limit external validity. However, the CBLs scenarios were real cases that represented the most common categories of adverse patient events.¹⁷ The response rate was low and data was missing for some participants who did not answer all the multiple choice questions, which may limit the sensitivity of our analyses. The web-based technology developed for this system is not available to other institutions. Our model did not map perfectly to Kember's in that we observed only a distinction between Habitual Action and the remaining items. Yet, previous research has emphasized the potential for factor instability and the importance of repeating factor analysis when using the same instrument in new educational settings, so this finding is not surprising.³¹ However, the basic contribution of this study is a validated method for measuring faculty physicians' reflections on adverse patient events, which could be utilized in future research and replicated elsewhere using either paper or electronic formats. Finally, study participants completed several cases apiece, which could be considered clustered data, which may limit our interpretation of the factor analysis.

In summary, reflection on adverse events is a crucial step in practice improvement.¹³ We describe what may be the first validated method for measuring degrees of faculty physicians' reflections on adverse patient events. This method will aid future research to compare quality outcomes among low versus high-reflecting physicians. We also found that reflection is enhanced by case material that is relevant and generalizable, which should be useful information when developing QI curricula for faculty physicians, as such curricula should strive to use actual, as opposed to hypothetical, case examples. The next challenge will be to

determine how to stimulate reflection among practicing physicians in order to improve the quality of healthcare.

Acknowledgments: None.

Funding: None.

Conflict of Interest: None disclosed.

Corresponding Author: Christopher M. Wittich, MD, PharmD; Department of Internal Medicine, Division of General Internal Medicine, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA (e-mail: Wittich.Christopher@mayo.edu).

REFERENCES

1. **Kohn LT, Corrigan JM, Donaldson MS.** To err is Human: Building a Safer Health System. Washington, DC: National Academy Press; 2000.
2. Joint Commission: Sentinel Event. Available at: <http://www.jointcommission.org/sentinelevents/>. Accessed October 7, 2010.
3. **Rachel MM, Stewart MW.** Establishing a mortality review process. *J Nurs Care Qual.* 2009;24(3):211–222.
4. **Szostek JH, Wieland ML, Loertscher LL, et al.** A systems audit increases awareness of systems-based practice at morbidity and mortality conference. *Am J Med.* 2010;123(7):663–668.
5. **Kerfoot BP, Conlin PR, Trivison T, McMahon GT.** Web-based education in systems-based practice: a randomized trial. *Arch Intern Med.* 2007;167(4):361–366.
6. **Frush K, Hohenhaus S, Luo X, Gerardi M, Wiebe RA.** Evaluation of a Web-based education program on reducing medication dosing error: a multicenter, randomized controlled trial. *Pediatr Emerg Care.* 2006;22(1):62–70.
7. **Stewart M, Marshall JN, Ostbye T, et al.** Effectiveness of case-based on-line learning of evidence-based practice guidelines. *Fam Med.* 2005;37(2):131–138.
8. **Peters A, Kimura J, Ladden M, March E, Moore G.** A self-instructional model to teach systems-based practice and practice-based learning and improvement. *J Gen Intern Med.* 2008;23(7):931–936.
9. **Kirkpatrick D.** Revisiting Kirkpatrick's four-level model. *Train Dev.* 1996;50:54–59.
10. **Beckman TJ, Cook DA.** Developing scholarly projects in education: a primer for medical teachers. *Med Teach.* 2007;29(2–3):210–218.
11. **Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM.** Association between funding and quality of published medical education research. *JAMA.* 2007;298(9):1002–1009.
12. **Shea J.** Mind the gap: some reasons why medical education research is different from health services research. *Med Educ.* 2001;35:319–320.
13. **Wittich CM, Reed DA, McDonald FS, Varkey P, Beckman TJ.** Transformative learning: a framework using critical reflection to link the improvement competencies in graduate medical education. *Academic Medicine.* 2010;85:1790–1793.
14. **Sanders J.** The use of reflection in medical education: AMEE Guide No. 44. *Med Teach.* 2009;31(8):685–695.
15. **Kember D, Leung D, Jones A, et al.** Development of a questionnaire to measure the level of reflective thinking. *Assess Eval High Educ.* 2000;25(4):381–395.
16. **Mann K, Gordon J, MacLeod A.** Reflection and reflective practice in health professions education: a systematic review. *Adv Health Sci Educ.* 2009;14:595–621.
17. **de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA.** The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care.* 2008;17:216–223.
18. **Boud D, Keogh R, Walker D.** Reflection: Turning Experience into Learning. London: Kogan Page; 1985.
19. **DeVellis RF.** Scale Development: Theory and Applications. Newbury Park, CA: Sage; 1991.
20. **Beckman TJ.** Determining the validity and reliability of clinical assessment scores. In: Henderson M, ed. *A Textbook for Internal Medicine Education Programs.* Washington, D.C.: Association of Program Directors in Internal Medicine and Association of Specialty Professors; 2007:139–146.
21. **Schon DA.** Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions. San Francisco: Jossey-Bass; 1987.

22. **Wittich CM, Beckman TJ, Drefahl MM, et al.** Validation of a method to measure resident doctors' reflections on quality improvement. *Med Educ.* 2010;44:248–255.
23. **Boenink AM, Oderwald AK, de Jonge P, van Tilburg W, Smal JA.** Assessing student reflection in medical practice. The development of an observer-rated instrument: reliability, validity and initial experiences. *Med Educ.* 2004;38:368–377.
24. **Roberts C, Stark P.** Readiness for self-directed change in professional behaviours: factorial validation of the Self-reflection and Insight Scale. *Med Educ.* 2008;42:1054–1063.
25. **Messick S.** Validity. In: Linn RL, ed. *Educational Measurement*. Phoenix, AZ: Onyx Press; 1993.
26. **Messick S.** Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50(9):741–749.
27. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
28. **Downing SM.** Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830–837.
29. **Beckman TJ, Cook DA, Mandrekar JN.** What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med.* 2005;20(12):1159–1164.
30. **Cook DA, Beckman TJ.** Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119(2):166 e167-116.
31. **Beckman TJ, Cook DA, Mandrekar JN.** Factor instability of clinical teaching assessment scores among general internists and cardiologists. *Med Educ.* 2006;40(12):1209–1216.