Exploring Validity Evidence Associated With Questionnaire-Based Tools for Assessing the Professional Performance of Physicians

Exploring Validity Evidence Associated With Questionnaire-Based Tools for Assessing the Professional Performance of Physicians: A Systematic Review

Mirja W. van der Meulen, MSc, Alina Smirnova, MD, PhD, Sylvia Heeneman, MD, PhD, Mirjam G.A. oude Egbrink, MD, PhD, Cees P.M. van der Vleuten, PhD, and Kiki M.J.M.H. Lombarts, PhD

Abstract

Purpose
To collect and examine—using an argument-based validity approach—validity evidence of questionnaire-based tools used to assess physicians' clinical, teaching, and research performance.

Method
In October 2016, the authors conducted a systematic search of the literature seeking articles about questionnaire-based tools for assessing physicians' professional performance published from inception to October 2016. They included studies reporting on the validity evidence of tools used to assess physicians' clinical, teaching, and research performance. Using Kane's validity framework, they conducted data extraction based on four inferences in the validity argument: scoring, generalization, extrapolation, and implications.

Results
They included 46 articles on 15 tools assessing clinical performance and 72 articles on 38 tools assessing teaching performance. They found no studies on research performance tools. Only 12 of the tools (23%) gathered evidence on all four components of Kane's validity argument. Validity evidence focused mostly on generalization and extrapolation inferences. Scoring evidence showed mixed results. Evidence on implications was generally missing.

Conclusions
Based on the argument-based approach to validity, not all questionnaire-based tools seem to support their intended use. Evidence concerning implications of questionnaire-based tools is mostly lacking, thus weakening the argument to use these tools for formative and, especially, for summative assessments of physicians' clinical and teaching performance. More research on implications is needed to strengthen the argument and to provide support for decisions based on these tools, particularly for high-stakes, summative decisions. To meaningfully assess academic physicians in their tripartite role as doctor, teacher, and researcher, additional assessment tools are needed.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Mirja W. van der Meulen, Professional Performance Research Group, Academic Medical Center, University of Amsterdam, Meibergdreef 9, PO Box 22700, 1100 DD Amsterdam, The Netherlands; telephone +31 (0) 20-566-1274; email: m.w.vandermeulen@amc.uva.nl or m.vandermeulen@maastrichtuniversity.nl; Twitter: @Mirja_vd_Meulen.

Acad Med. 2019;94:1384–1397.
First published online April 23, 2019
doi: 10.1097/ACM.0000000000002767

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A677.

Physicians' professional performance consists of activities done to fulfill their tripartite role as clinicians, teachers, and researchers.[1] To support them in their ongoing professional development, assessing performance in these activity areas is of vital importance.[2] Workplace-based assessment methods enable the academic medicine community to assess professional performance, and thus give insight into the actual performance of physicians in daily practice.[3] Questionnaire-based tools serve as a means to collect valuable information about physicians' professional performance in a feasible and comprehensive way from those who can and do observe them in their daily workplace.[4,5] Multisource feedback tools are an example of questionnaire-based tools; they consist of questionnaires with multiple items and rating scales used to collect and assess performance information.

Although a plethora of questionnaire-based tools designed to get insight into physicians' capabilities for both clinical practice and teaching medicine are available, ensuring that these tools generate trustworthy data is crucial for providing physicians with relevant performance feedback and/or making sound decisions about remediation or promotion. Thus far, investigators have gathered and meticulously investigated the validity evidence of these tools yet failed to prioritize among the different sources of validity evidence.[4,6–10] For the validation process, understanding and prioritizing among these sources of validity evidence is crucial; tools used for formative purposes require different sources of evidence than tools used for summative purposes. Questionnaire-based tools for summative decisions inevitably need more validity evidence in general, and especially more evidence related to the implications or consequences of a decision. Ultimately, validity is about collecting evidence to defend the decision made based on the data resulting from the tool.[11] This need for differentiation and prioritization of validity evidence is now recognized as central to the debate regarding the validity of assessing physicians' professional performance.[12]

A state-of-the art approach to validity, articulated by Kane,[13] prioritizes among different sources of evidence and indicates how their priority varies for different assessment tools and purposes. The validation process can be seen as a structured validity argument consisting of multiple components (or inferences)—namely, scoring, generalization, extrapolation, and implications (see Method for more detailed explanation). To make a strong argument, evidence

regarding all components is necessary. Further, validity evidence on these components should not be examined in isolation from one another; the validity argument is a chain of inferences, and the strength of the argument is most influenced by the weakest link in the chain.[14]

Through this systematic review, we have collected and examined available validity evidence of published questionnaire-based tools used to assess physicians' professional performance. Applying Kane's framework[13] to the ongoing validity debate of questionnaire-based tools, we believe, opens up new possibilities to reframe the study of the validity of these tools. Our research question is, *How strong is the validity argument to support the use of and decisions resulting from questionnaire-based tools to assess physicians' clinical, teaching, and research performance?*

Method

Before conducting the review, we agreed on eligibility criteria, search strategy, study selection, data extraction, and study quality assessment. We performed our review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards.[15]

Data sources and search strategy

We conducted a systematic search of the literature on October 5, 2016, seeking articles on questionnaire-based tools for assessing physicians, published from inception to October 2016. We searched the following electronic databases: PubMed, ERIC, PsycINFO, and Web of Sciences. We limited our search to English-language, peer-reviewed journals. A clinical librarian assisted with the development of our search strategy and helped to specify keywords. We used both free-text and MeSH (MEDLINE) or thesaurus (Embase and PsycINFO) terms to indicate study topic, aim of the questionnaire-based tool, type of performance being assessed, how physicians were assessed, and the subjects of assessment (see our complete search strategy in Supplemental Digital Appendix 1, available at http://links.lww.com/ACADMED/A677). In addition, we searched the reference lists of included studies to find additional eligible studies.

Eligibility criteria

We considered studies eligible if they reported on a questionnaire-based tool for assessing physicians' clinical, teaching, and/or research performance. Inclusion criteria were as follows: (1) the article described one or more questionnaire-based tools that relies on colleagues, coworkers, residents, and/or patients as respondents to assess physicians' performance in practice, (2) the article reported on the questionnaire tool or its design, and (3) the article provided information about the validation process. Studies were excluded if (1) the tool was used to assess medical students, residents, and/or nonphysician health professions (e.g., nurses) and/or if (2) the tool was based solely on patients' responses.

Study selection

One author (M.W.vdM.) performed the initial search, which was duplicated by a clinical librarian. Subsequently, this author (M.W.vdM.) screened both the title and the abstract of all the titles found in the initial search. If the titles did not provide sufficient information, this author read the abstract and, at this point, excluded studies whose titles/abstracts did not mention physicians, assessment of performance, questionnaire-based tools, and information about validity. After this screening, two authors (M.W.vdM. and A.S.) independently reviewed, respectively, one-half of the remaining titles and abstracts for inclusion using the same criteria. Next, these two authors (M.W.vdM. and A.S.) each independently reviewed the full texts of all the remaining articles, again using the inclusion criteria described above. Discrepancies were resolved by discussion with a third author (K.M.J.M.H.L.) until the three achieved 100% agreement.

Data extraction and validity quality assessment

Once articles were identified for inclusion, two authors (M.W.vdM. and A.S.) extracted data from 20 studies collaboratively, and then they extracted data from the remaining studies individually. The data extracted from the studies comprised the following:

1. name of the tool (if no specific name was provided, the generic term "questionnaire-based tool" was used),
2. specialty of physician participants,
3. number of physicians assessed,
4. number and type of assessors,
5. country of origin,
6. number and type of items in the tool, and
7. feasibility of the tool (duration and costs, platform used, number of assessors needed).

Next, the two authors extracted data about the validation process of each tool based on Kane's validity approach. Kane takes an argument-based approach to examining validity; his approach consists of two types of arguments: (1) the interpretation/use argument and (2) the validity argument. The validation process starts with naming the claims that are being made in a proposed interpretation or use (the interpretation/use argument) for a given tool, and then moves on to evaluating these claims (the validity argument).[16] Thus, we sought data about the evidence that the authors of the included studies provided to support their claims.

First, we extracted the authors' interpretation of the assessment data/test scores and their proposed use of the tool. For example, a statement such as "A score of 8 out of 10 indicates good performance, and anyone scoring higher than 8 should be given promotion" indicates an interpretation and proposed use. Without the interpretation of data, validation is useless because the framework for the validity argument is not stated, and thus no specific evidence can be collected.[13]

Second, we extracted information on the validity argument for each tool. The validity argument consists of four components—scoring, generalization, extrapolation, and implications—which together create a coherent chain of inferences to support the intended interpretations and uses.[13]

Scoring. The scoring component of the argument requires information about how the assessment data were collected, recorded, and scored.[17] For questionnaire-based tools, evidence about the scoring component should contain information about the following:

- how the items were developed,

- whether the assessors had ample opportunity to observe the physician (so they can score the physician fairly/adequately),
- how assessors were sampled (are they selected by the physicians themselves, or by a third party?),
- if assessors assessed the physicians voluntarily and anonymously, and
- whether assessors received sufficient explanation on how to score items.

That is, evidence on questionnaire-based tools addresses the question of whether the scoring criteria were appropriate and correctly applied: Were the items, scales, and raters appropriate?

Generalization. The generalization component focuses on the link between the observed sample of performance and the wider domain of all possible performances in the assessment setting. Evidence for this component involves classical test theory or generalizability theory and answers the question, "Do these specific items and raters used in this particular assessment setting generalize to other items and raters in this setting?"

Extrapolation. Extrapolation is about whether the observations made are linked to the real-world activity of interest. The focus of this component is on collecting evidence showing the relationship between the construct of interest and the scores obtained. The intent is to answer the question, "Can we extrapolate the scores seen in this assessment context to outcomes in other assessment contexts or in real clinical performance?" Evidence includes factor analyses, investigations of desired relationships between scores and other measures, and identifying expected performance level differences.[17]

Implications. The last component of the validity argument is about the implications—that is, what the consequences of the assessment are for the physician, other stakeholders, and society at large.[11] Consequences can result either from the use of assessment data or from the mere act of assessing the physician. Evidence about this inference could most straightforwardly emanate from offering the assessment (and the ensuing judgment and intervention [e.g., promotion or remediation]) to some physicians, but not to others, and then comparing the consequences and impact that follow.[11]

To determine the quality of the validity evidence per component, we adapted the quality checklist used by Beckman and colleagues[7] to fit the argument-based validity framework (see Table 1). The original checklist[7] was based on

Table 1

Criteria for Rating Validity Evidence of Data From Questionnaire-Based Assessment Tools, Based on Beckman and Colleagues' (2005)[7] Rating Criteria[a]

Component category	Evidence category	Rating score	Rating criteria
Scoring	Item development	0	No discussion of instrument content (includes simply listing items without justification)[b]
		1	Discussion but no data (simply stating items were properly developed)[b]
		2	Listing items with little or no reference to a theoretical basis, or a poorly defined process for creating and reviewing items[b]
		3	Well-defined process for developing instrument content, including both an explicit theoretical/conceptual basis for instrument items and systematic item review by experts. Alternatively, reference to a prior study on an assessment instrument that meets these criteria[b]
	Raters	0	No discussion[c]
		1	Discussion but no data. Merely disclosing response rates or numbers of respondents or type of selection does not constitute evidence[b]
		2	Discussion and/or minimal data about *how raters were appropriate, or able to assess, or discussion of biases*[c]
		3	Multiple sources of supportive data *on demonstrating appropriateness of raters (no biases found with bias study, evidence that raters were able to observe, unable to assess option/rate discussed and followed up)*[c]
	Scores and scales	0	No discussion of the scoring process, scale use, or guidelines[d]
		1	Discussion but no data (only description of guidelines and/or exemplary behavior given)[d]
		2	Minimal data: only guidelines and descriptives on scores given, yet no follow-up[d]
		3	Multiple sources of supportive evidence; guidelines given, exemplary behavior and follow-up on non-normal scores[d]
Generalization	Reliability	0	No discussion[d]
		1	Discussion but no data[d]
		2	Minimal data: only Cronbach alpha reported[d]
		3	Data: Cronbach alpha reported and higher than 0.80 for whole instrument[d]
	G study	0	No discussion[d]
		1	Discussion but no data[d]
		2	G study performed, yet reported G coefficients < 0.80 or only number of raters or standard errors stated[d]
		3	G study performed, with reported G coefficients > 0.80 and number of raters stated[d]

(*Table continues*)

Table 1
(Continued)

Component category	Evidence category	Rating score	Rating criteria
Extrapolation	Constructs	0	No discussion[b]
		1	Discussion but no data[b]
		2	*(Exploratory)* Factor analysis incompletely confirming anticipated data structure, or acceptable reliability with a single measure[c]
		3	*(Confirmatory)* Factor analysis confirming anticipated data structure, or multiple measures of reliability[c]
	Performances	0	No discussion[b]
		1	Discussion but no data[b]
		2	Correlation of assessment scores to outcomes with minimal theoretical importance, or unanticipated score correlations[b]
		3	Correlation (convergence) or no correlation (divergence) between assessment scores and theoretically predicted outcomes or measures of the same *or different construct*. Such evidence will usually be integral to the study design, and anticipated a priori[c]
Implications	Intended outcomes	0	No discussion[b]
		1	Discussion but no data: *Speculation on potential performance improvement does not constitute evidence, nor does stating the proportion of respondents intended to improve*[c]
		2	Minimal data provided: description of performance change (self-identified, likelihood of change, or other scores)[d]
		3	Multiple data provided: changes in scores, changes in other measurement, objective impact in health care[d]
	Unintended outcomes	0	No discussion[b]
		1	Discussion but no data. Simply discussing the consequences of assessment (e.g., data regarding usefulness or faculty approval) without linking this to validity does not constitute evidence. *Speculation on potential applications of the assessment does not constitute evidence*[c]
		2	Description of consequences of assessment that could conceivably impact the validity of score interpretations (although these impacts are not explicitly identified by the authors). *Discussion of nonappropriate group differences with data, but no follow-up*[c]
		3	Description of consequences of assessment that clearly impact the validity of score interpretations, as supported by data and convincingly argued by the authors. Such evidence will usually be integral to the study design, and anticipated a priori[b]

[a]These rating criteria were based on and adapted from Beckman et al.[7] Adapted table printed with permission of Springer Nature: Springer Nature. Journal of General Internal Medicine. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? 2005;20. The footnotes below indicate whether and how the criteria were adapted.
[b]Copied from Beckman et al.
[c]Adaptions indicated in italics.
[d]Criteria not adapted from Beckman et al.[7]

operational definitions of the five sources of validity evidence per the *Standards for Educational and Psychological Testing* published by the American Psychological Association and the American Education Research Association.[18] Two authors (M.W.vdM. and A.S.) scored the validity evidence, based on the following format:

0 = no discussion of this source of validity evidence and/or no data presented;

1 = discussion of this source of validity evidence, but no data presented, or data failed to support the validity of instrument scores;

2 = data for this source weakly support the validity of score interpretations; and

3 = data for this source strongly support the validity of score interpretations.

Data synthesis and analysis

We have presented our findings descriptively in text, tables, and figures to give a systematic overview of the validity evidence for the use of questionnaire-based tools. We have summarized the strength of the validity argument by averaging the quality rating scores given to the tools—both (1) per component and for the complete argument and (2) for all tools and for only tools that provided evidence. To evaluate the validity argument, we assumed that questionnaire-based tools for assessing physicians could have two uses—formative or summative—and we weighted the evidence accordingly. We weighted the evidence, based on the literature on assessment and the argument-based approach to validity,[17] setting an arbitrary cutoff score of 1.50 for all components for formative purposes, and, because higher-stakes claims require more evidence, a higher cutoff score of 1.80 for summative purposes.

Results

Number of studies and tools

From the 8,533 initial hits our database and hand search garnered, we identified 46 relevant studies[3,19–63] describing tools designed for assessing physicians' clinical performance and 72 studies designed for assessing their teaching performance.[64–135]

We found no tools designed to assess physicians' research performance. From the 46 articles on clinical performance tools, we identified 15 unique tools, and from the 72 articles on teaching performance, we identified 38 unique tools. For details regarding the selection process, see Figure 1, and for details about the included studies' settings, assessors, and subjects, see Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/A677.

The validity argument for questionnaire-based assessment tools

Examining the complete validity argument requires considering whether evidence has been collected on all four components of the argument (scoring, generalization, extrapolation, and implications). Five clinical performance tools gathered evidence on all components of the validity argument.[19–31,34–39,42–49,53,55,57–61] The remaining tools most often neglected evidence for intended implications. Seven teaching performance tools collected evidence on all components of the argument.[74,78,83–85,91,92,96,98,99,101,103,106,108,109,111,113,115,117,118,120–123,128,131–134] Thus, in total, only 12 (23%) of all 53 tools gathered evidence on all four components of Kane's validity argument.

Below, we describe the results within each component of the validity argument, or chain of inferences, separately: first, for clinical performance tools and, second, for teaching performance tools. See Table 2, Figure 2, and Table 3 for a comprehensive overview of the strength of the validity argument for the questionnaire-based tools.

Evaluating the inferences of the validity argument

Supplemental Digital Appendix 3, available at http://links.lww.com/ACADMED/A677, summarizes the results of the modified quality checklist applied to the various components of the validity argument for each type of performance tool, and we have described the results for each of the components of the validity argument

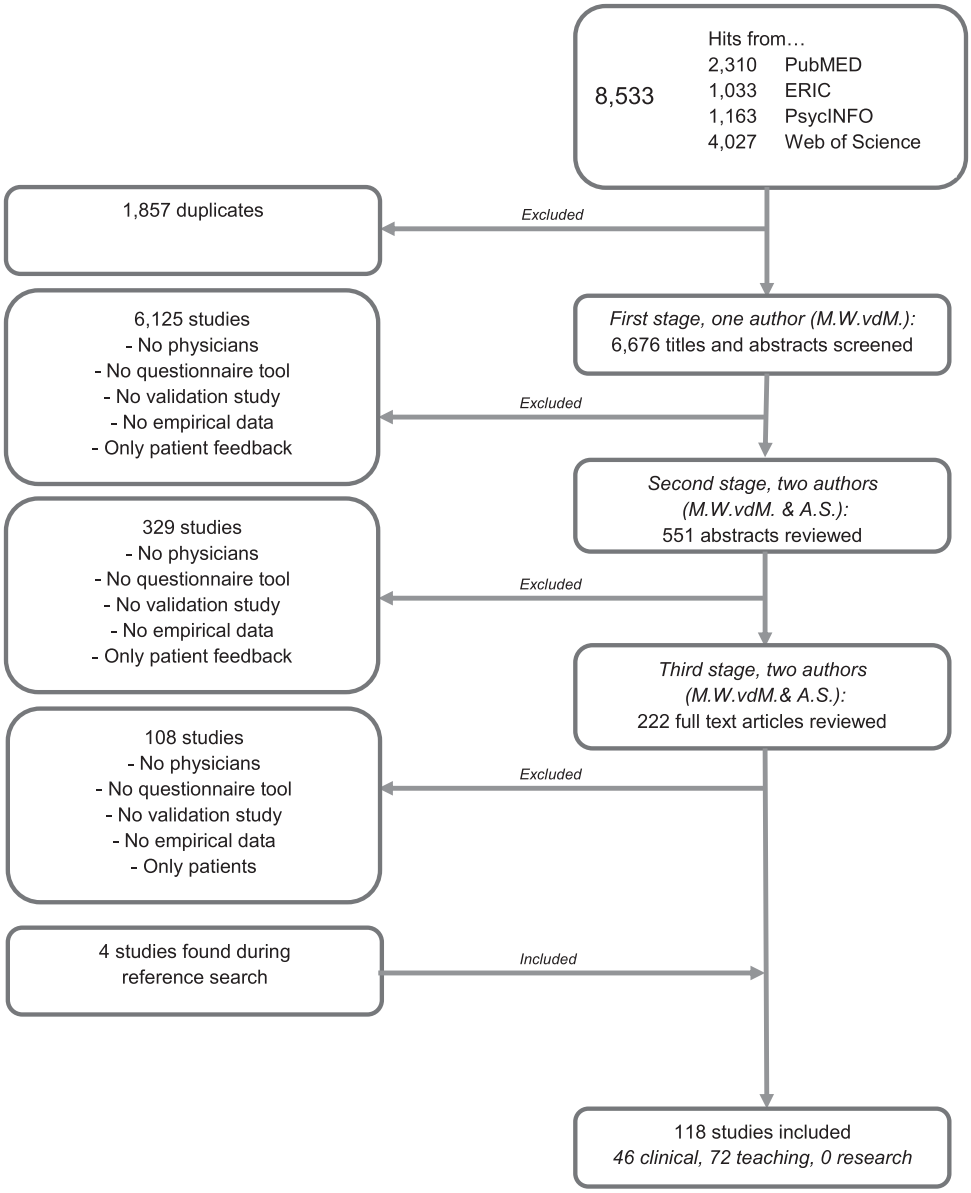

Figure 1 Flowchart of the study selection and review process for a systematic review of the literature on questionnaire-based assessment tools for physicians' clinical, teaching, and research performance, published 1966–October 2016.

Table 2

Summary of Validity Evidence of the 118 Studies on Questionnaire-Based Assessment Tools for Physicians' Clinical and Teaching Performance Included in a Systematic Analysis of the Literature Published 1966–October 2016

			Clinical performance tools[a]		Teaching performance tools[a]	
Component category	Evidence category	Data category	All tools, n = 15	Only tools that provide evidence[b]	All tools, n = 38	Only tools that provide evidence[b]
Scoring	Item development	Mean score[c]	2.27	2.83	1.65	2.52
		SD of score	1.18	0.37	1.32	0.70
		No. (%) of tools providing evidence	12 (80%)	12 (80%)	25 (65%)	25 (65%)
		Minimum–maximum score of included studies	0–3	2–3	0–3	1–3
	Raters	Mean score[c]	1.28	1.36	0.92	1.35
		SD of score	0.57	0.48	0.74	0.48
		No. (%) of tools providing evidence	14 (93%)	14 (93%)	26 (68%)	26 (68%)
		Minimum–maximum score of included studies	0–2	1–2	0–2	1–2
	Scores and scales	Mean score[c]	1.13	1.31	0.37	1.17
		SD	0.72	0.61	0.58	0.37
		No. (%) of tools providing evidence	13 (87%)	13 (87%)	12 (32%)	12 (32%)
		Minimum–maximum score of included studies	0–3	1–3	0–2	1–2
Generalization	Reliability	Mean score[c]	1.33	2.86	1.74	2.87
		SD of score	1.45	0.35	1.43	0.34
		No. (%) of tools providing evidence	7 (47%)	7 (47%)	23 (61%)	23 (61%)
		Minimum–maximum score of included studies	0–3	2–3	0–3	2–3
	G study	Mean score[c]	1.47	2.75	0.89	2.83
		SD of score	1.40	0.43	1.33	0.37
		No. (%) of tools providing evidence	8 (53%)	8 (53%)	12 (32%)	12 (32%)
		Minimum–maximum score of included studies	0–3	2–3	0–3	2–3
Extrapolation	Constructs	Mean score[c]	1.33	2.22	1.03	2.29
		SD score	1.14	0.42	1.18	0.46
		No. (%) of tools providing evidence	9 (60%)	9 (60%)	17 (45%)	17 (45%)
		Minimum–maximum score of included studies	0–3	2–3	0–3	2–3
	Performance	Mean score[c]	1.13	2.13	1.53	2.76
		SD score	1.09	0.33	1.41	0.43
		No. (%) of tools providing evidence	8 (53%)	8 (53%)	21 (55%)	21 (55%)
		Minimum–maximum score of included studies	0–3	2–3	0–3	2–3
Implications	Intended	Mean score[c]	0.67	1.67	0.53	2
		SD score	0.87	0.47	0.88	0
		No. (%) of tools providing evidence	6 (40%)	6 (40%)	10 (26%)	10 (26%)
		Minimum–maximum score of included studies	0–2	1–2	0–2	2–2
	Unintended	Mean score[c]	0.53	2	0.21	2
		SD score	0.88	0	0.61	0
		No. (%) of tools providing evidence	4 (26%)	4 (26%)	4 (11%)	4 (11%)
		Minimum–maximum score of included studies	0–2	2–2	0–2	2–2

Abbreviation: SD indicates standard deviation.

[a]The columns under the "Clinical performance tools" and "Teaching performance tools" headings indicate whether all tools were taken into account, or whether only the tools that provided evidence on that particular inference were taken into account.

[b]The number of tools that provided evidence varied depending on the evidence category.

[c]Possible mean scores range from 0 to 3, with 0 being the lowest score and 3 the highest, indicating, respectively, no evidence of validity to high quality of evidence.

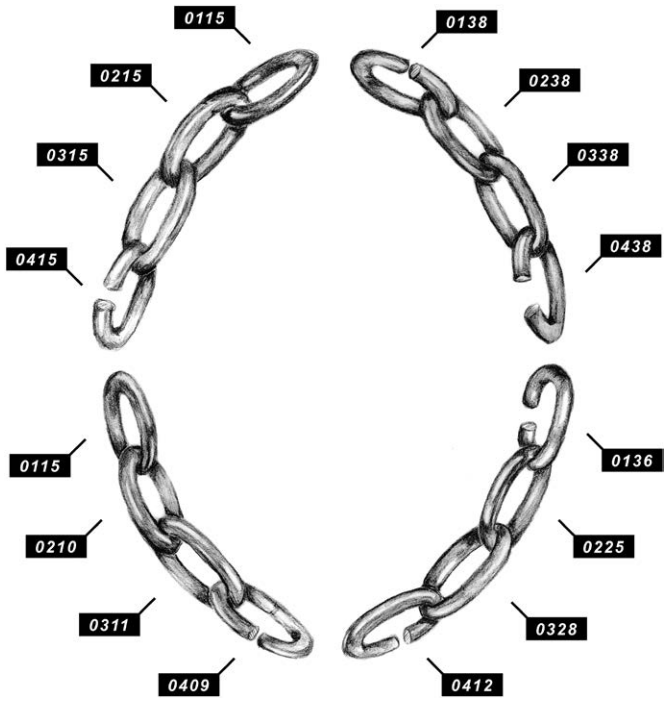

Figure 2 The strength of the validity argument for assessments of physicians' clinical and teaching performance, depicted as a chain of inferences[137] for the 53 questionnaire-based assessment tools included in a systematic analysis of the literature published 1966–October 2016. In this chain, every inference of the validity argument is represented as a link in the chain. The numbers on the links are paired with the strength of the validity which can be found in Table 3. Each chain number is constructed from two digits: The first two digits represent the four components—01 *scoring*, 02 *generalization*, 03 *extrapolation*, and 04 *implications*—and the last two digits represent the number of tools. Drawing: Mirja van der Meulen, Amsterdam, The Netherlands. Graphical design: Turkenburg Media, Haarlem, The Netherlands.

in detail below. We provide specific examples either to show best practices of validation processes or to show conflicting results in the validity evidence of questionnaire-based tools.

Evidence for scoring. Overall, tools for clinical performance assessment gathered evidence on, primarily, the appropriateness of item development, whereas the evidence on the appropriateness of raters and scale use was mixed. Across the 46 articles describing all 15 clinical performance tools, we calculated an average evidence score of 1.55 (standard deviation [SD] = 0.58). Teaching performance tools gathered less evidence on the scoring component: Across all 72 articles describing the teaching performance tools, we detected an average evidence score of 0.98 (SD = 0.59); however, the score was a bit higher—1.04 (SD = 0.57)—when we excluded tools that did not gather any evidence on the scoring inference.

Item development. Investigation into the appropriateness of the items revealed that 41 studies developed clinical performance tools based on a theoretical framework, peer-reviewed literature, other documents, other preexisting tools, or expert opinions.[3,19–31,33–40,42–45,47–49,51–61,63] For the teaching tools, the scoring inference for item development seems to be overlooked by most authors. Studies of 21 tools do not or only poorly disclose how tools were developed regarding the items, scoring, or scales.[64,67,68,74–76,78,82,84,87,90–92,97,98,100,104,110,111,114,115,125,127,130] Studies on the remaining 17 tools disclosed how items were developed based on a theoretical framework, peer-reviewed literature, other documents, other validated tools, or expert opinions.[65,66,69,72,73,77,79,81,83,85,88,89,93–96,99,101–103,105–109,112,113,116–124,126,128,129,131–135]

Raters. Most of the identified studies did not provide validity evidence for the appropriateness of raters. Studies on clinical performance tools provided limited information about the impact of rater selection on assessment scores. Almost all studies on clinical performance assessment tools[3,19–32,34–49,53–55,57–62] used physician self-selected raters—based on the studies of Ramsey and colleagues which indicated that self-selection had a negligible effect on scores.[19–23] However, one study investigated the method the National Clinical Assessment Service (NCAS) used to select raters who assessed referred physicians.[52] This study found that, for physicians in potential difficulty (NCAS referred), self-selected raters gave significantly higher scores compared with raters who were selected by the referring body. That is, when a physician selected his/her own raters, especially in a high-stakes setting, resulting scores were more positive than results from raters who were not selected by the physician. For tools used to assess teaching, information on rater selection was mostly lacking. In fact, only two teaching assessment tools stated that raters could self-select faculty assessors, and one tool used a randomization process to select raters.[95,96,98,101–103,106,108,109,117,118,120,122,123,128,131–133] Whether raters had ample opportunity to observe the physician was acknowledged by only three clinical assessment tools, although almost every tool included an "unable to assess" option for raters.[19–21,23,27,56,63] For teaching performance tools, over a third of the tools (n = 28) did not mention whether raters could select "unable to assess."[64–66,69,70,74–87,89–95,97–100,102,104,111,114–116,119,121,125,127,129,130,134]

Scores and scales. Four studies on clinical performance tools do not report the distribution of ratings,[32,33,51,56] and the 42 that do all indicate that scores were highly skewed to favorable impressions of physicians' clinical performance. It is unclear whether these generally favorable scores indicate genuinely excellent performance or colleagues' reluctance to identify below-average performance, especially within high-stakes settings. The study of Archer and McAvoy[52] illuminates this phenomenon; negatively skewed distributions of ratings were found for NCAS-referred doctors who self-selected their assessors, whereas a more normal distribution was found for these doctors when they were assessed by referring-body-selected raters. Twelve articles on tools assessing teaching performance reported descriptive statistics of the scale scores, yet not one examined whether, and if so, how and why, scores were skewed.[66,71,73,75,79,89,91,92,94,96,97,100,101,103,104,106–109,112,113,116–118,120,122,123,127–133,135]

Table 3

The Strength of Each Link of the Validity Argument of Assessments of Physicians' Clinical and Teaching Performance, Depicted as a Chain of Inferences[137] for the 53 Questionnaire-Based Assessment Tools Included in a Systematic Analysis of the Literature Published 1966–October 2016

Chain number[a]	Mean (SD) validity evidence score	No. (%[b]) of tools with evidence	Minimum and maximum score	Type of performance tool
0138	0.98 (0.59)	36 (95)	0–2.33	Teaching
0238	1.32 (1.15)	25 (66)	0–3	Teaching
0338	1.28 (0.93)	28 (74)	0–3	Teaching
0438	0.37 (0.58)	12 (32)	0–2	Teaching
0136	1.04 (0.57)	36 (95)	0.33–2.33	Teaching
0225	2 (0.80)	25 (66)	1–3	Teaching
0328	1.73 (0.62)	28 (74)	1–3	Teaching
0412	1.17 (0.37)	12 (32)	1–2	Teaching
0115	1.55 (0.58)	15 (100)	0.67–2.67	Clinical
0210	2.10 (0.74)	10 (67)	1–3	Clinical
0311	1.68 (0.57)	11 (73)	1–2.50	Clinical
0409	1 (0.41)	9 (60)	0.50–2	Clinical
0115	1.55 (0.58)	15 (100)	0.67–2.67	Clinical
0215	1.40 (1.16)	10 (67)	0–3	Clinical
0315	1.23 (0.89)	11 (73)	0–2.50	Clinical
0415	0.60 (0.58)	9 (60)	0–2	Clinical

[a]The chain numbers are constructed from two digits: The first two digits represent the four components—01 *scoring*, 02 *generalization*, 03 *extrapolation*, and 04 *implications*—and the last two digits represent the number of tools with evidence. See also Figure 2.
[b]The percentage represents the portion of tools with evidence out of, respectively, the 38 total teaching tools and the 15 total teaching tools.

Evidence for generalization. On average, across the studies reporting on clinical assessment tools, we calculated a score of 1.40 (SD = 1.16), and across the studies of teaching performance tools, we calculated a score of 1.32 (SD = 1.15). When we excluded the tools that did not provide evidence on this component, we calculated a mean score of 2.10 (SD = 0.74) and 2.00 (SD = 0.80) for, respectively, clinical and teaching assessment tools.

Reliability. Review of the research indicates that most clinical and teaching tools provide evidence of internal consistency; Cronbach α is generally higher than 0.80 both for subscale scores and for overall scores.[24–26,28–31,34–39,41–45,47–50,53,55,57–61,63,67,72–74,78,81–85,87,91–96,98,101–109,112,113,116–118,120–126,128–135]

Generalizability. Data from the studies that investigated the generalizability of clinical performance assessment tools suggest that, on average, 10 coworkers would be sufficient to produce a generalizability coefficient higher than 0.80.[3,19–31,34–38,42–47,49,50,54,55,61,63] Data from the studies on 10 teaching tools indicate that, on average, ratings from 13 learners are necessary for reliable estimates.[71,92,96,102,107,109,113,116,124,128,130]

Evidence for extrapolation. Across the 46 articles on clinical performance assessment tools, the average extrapolation inference score was 1.23 (SD = 0.89); however, that score rose to 1.68 (SD = 0.57) when we excluded tools that did not provide evidence on extrapolation. Across the articles about the teaching performance assessment tools, the average extrapolation score was 1.28 (SD = 0.93), but higher—1.73 (SD = 0.62)—when we included only the tools that provided evidence.

Link to performances and group differences. Three studies on clinical performance assessment tools related test scores to other variables of interest. Ramsey and colleagues[20] found that internists who were rated highly by their associates also had high American Board of Internal Medicine licensure exam scores. A study on the General Medical Council (GMC; United Kingdom) colleague questionnaire (CQ) showed that the CQ scores were positively correlated with the Colleague Feedback Evaluation Tool, a similar tool that assesses physicians' clinical performance.[60] Another study indicated that the GMC CQ scores positively correlated with the number of positive comments provided by colleagues.[48] For tools assessing teaching, one study found that comments were more likely for negative evaluations, and the length of these comments correlated negatively with the assessment score: the more written feedback, the lower the score.[124] Receiving more positive comments also significantly and positively correlated to teaching scores.[117] Three studies tried to elucidate the relationship between teaching and clinical performance. Physician subgroups performing more than two major procedures per week at the hospital received higher ratings from students than those who did not.[67] McOwen and colleagues[92] found a significant and positive correlation between clinical excellence and ratings of teaching excellence given by residents. Finally, the study of Mourad and Redelmeier[87] reported no significant associations between teaching effectiveness scores and adverse patient outcomes.

One study scrutinized expected clinical performance level differences: Physicians who had indications of performance concerns received significantly lower scores than a volunteer sample of physicians, yet the effect sizes were small.[52] The results for tools assessing teaching performance by rank were conflicting: Professors had higher teaching scores in one study,[83] whereas another study showed no significant differences among academic ranks.[134] The findings of other studies on teaching assessment tools, however, did support the extrapolation inference: Backeris and colleagues[114] found that academic faculty received significantly higher teaching scores compared with clinical faculty. Additionally, a study on a teaching performance tool intended for emergency medicine (EM) faculty showed that EM-certified faculty received significantly higher scores than non-EM-certified faculty.[78] Furthermore, recently certified physicians, those who had attended a teacher training program, and those who spent more time teaching than seeing patients or conducting research all received high teaching scores.[108] Finally,

physicians who had been nominated as best teacher[93] or who had won a teaching award received higher teaching scores.[75]

Constructs. For clinical performance, 19 studies on 9 different tools showed that certain items were logically clustered in domains of performance with exploratory factor analyses.[21,23,24,30,31,33,35–37,39,41,42,44–47,50,58,63] Of these 19 studies, only 2 confirmed the found structure with a well-fitting confirmatory factor analysis.[23,44] These tools typically examined domains such as "Professionalism," "(Clinical) Competency," and "Collaboration." For teaching performance, 14 tools sought evidence by exploratory factor analysis,[65,68,72,73,85,91,93,96,100,103,104,106,109,124,126,128,130,131] and of these 14, only 2 sought further evidence through confirmatory factor analysis.[72,96,101,103,106,108,117,118,120,122,123,126,128,131–133] Investigators of 3 tools performed only a confirmatory factor analysis—not an a priori exploratory factor analysis.[102,111,113] Teaching tools most commonly measured performance domains such as "Clinical Teaching," "Interpersonal Skills," and "Learning Climate."

Evidence for implications. Across the 46 articles focused on clinical performance assessment, and the 72 articles on teaching assessment, the average implications evidence score was, respectively, 0.60 (SD = 0.58) and 0.37 (SD = 0.58). When we considered only the tools that provided evidence for implications, the average score became, respectively, 1.00 (SD = 0.41) and 1.17 (SD = 0.37).

For the clinical performance assessments, 11 studies reported self-identified or intended change of practice of assessed physicians.[25–28,43,44,49,51,59,61,62] Of these, 9 reported that more than half of the participants intended to make, or had already made, changes to their performance.[25–28,43,44,49,59,61] Interestingly, those physicians who felt that they performed better than their colleagues had rated them were less prone to make changes to their practice.[49] Violato and colleagues[44] investigated whether physicians' scores changed after a period of time and found a significant, yet small positive effect for physicians' mean aggregated scores. The lack of studies investigating the impact of clinical performance assessment on health care—the ultimate goal—is striking.

For teaching tools, seven studies investigated whether scores changed over time and showed an improvement in scores after one or several assessment periods.[65,70,84,98,115,121,133] One study found a significant change in scores after physicians received teacher training, and one study showed that after receiving the assessment feedback, faculty received significantly higher ratings over time.[70,121] Physicians who discussed their scores after the assessment had better subsequent scores compared both with those who did not discuss the feedback and with those who did not receive their scores.[65] A study on self-identified change showed that most physicians were positive about their improvement.[113] Another study identified that one factor negatively affecting intention to change is the experience of negative emotions in faculty themselves or recognizing negative emotions in others.[118]

Discussion

Main findings

We conducted this systematic review to collect and examine the validity evidence for questionnaire-based tools used to assess physicians' clinical, teaching, and research performance, for both formative and summative purposes. We identified a total of 15 questionnaire-based tools for physicians' clinical performance, 38 tools for physicians' teaching performance, and none for research performance. After reviewing the evidence through the four inferences of Kane's validity framework—scoring, generalization, extrapolation, and implications—our overall conclusion is that reasonable evidence supports the use of questionnaire-based tools to assess clinical performance for formative purposes, as the average scores were higher than 1.50 for tools that provided evidence. The arguments for using these tools to assess clinical performance for summative use, and for using them to assess teaching performance for either summative or formative use, lack crucial evidence in the implications component and thus should be used with caution. Furthermore, not all questionnaire-based tools seem to be supportive for their intended use.

Explanation of findings and suggestions for future research

In Kane's[13,16] argument-based approach to validation, evidence regarding all four components together creates a coherent and complete chain of inferences to support the intended interpretations and uses of assessment tools. Using this chain metaphor, it follows that the chain of inferences is only as strong as its weakest link, and strong evidence for one component of an argument does not compensate for weaknesses in other components of the argument (Figure 2 and Table 3).[13] Our review shows that the generalization and extrapolation components have received sufficient attention from researchers, the scoring component shows conflicting results, and the evidence surrounding the implications component is mostly lacking. This lack constitutes a serious limitation to using these questionnaire-based tools, in particular for summative purposes. The few studies that included implications evidence focused only on self-identified improvement or changes in assessment scores after some period of time; thus, the existing implications evidence does not provide strong support for using questionnaire-based tools. When assessment tools are employed to ensure (minimum) performance levels (i.e., that physicians are competent clinicians or teachers), then more supporting evidence is needed. Filling the gap of implications evidence is, therefore, crucial when assessment tools are used for summative purposes. We acknowledge that collecting strong implications evidence is a difficult endeavor—necessitating procedures that provide data on the both the assessment itself and the ensuing judgments to specific physicians.[11] Nevertheless, filling this gap in implications evidence is crucial, and future investigators could consider experimental designs, use appropriate statistical models for observational designs (e.g., g-estimation), and/or collaborate with other research fields.[136] Especially today, given the recent developments in accountability and public transparency, the academic medicine community must strive for implications evidence, even though doing so is difficult in the vast and context-specific field of medical education.

Additionally, this review has provided some conflicting results regarding the scoring component of the argument, which also weakens the validity argument. Although the item development of most tools for assessing clinical performance was properly developed, we noted issues

about the appropriateness of raters and scales (i.e., the effect of the rater selection and the lack of research on the negative skewing of scale scores). Therefore, future research on the scoring component should address the effect of the type of selection of raters and the use of the scoring scales. A possible explanation to these findings is that most studies were based within the "construct-model validity" approach, the most dominant discourse of validity in the past.[137,138] None of the studies approached the collection of validity evidence with an argument-based approach, which could explain why these components of the argument have been overlooked: Authors were simply less aware of that type of evidence.

Interestingly, we found no questionnaire-based tools used to assess physicians' research performance. This lack may not be surprising given the citation metrics—h-index, plus the number of publications, grants, clinical trials, and awards/honors received—that are available to assess physicians' research performance.[139,140] Notably, however, a strict focus on these types of metrics does not provide insight into the full scope of research performance—and might even *de*crease research performance.[141] Hence, other assessment tools should be considered, such as questionnaire-based tools based on physician competency frameworks.[1,2]

Practical implications

Although we found no completely valid argument for the use of questionnaire-based tools for assessing physicians, we feel that the academic medicine community should not reject these tools as a whole. The notion that not one single type of tool is superior to another aligns with theories on assessment and evaluation.[142] Every tool in an assessment program has its own strengths, weaknesses, and purpose and should be regarded as just one imperfect tool designed for a specific end. Through this review, we have elucidated the strengths and weaknesses of questionnaire-based tools, thus providing a guide for those interested in setting up meaningful assessment programs for physicians. Currently, the strength of these tools lies within the generalization and extrapolation components of the argument. Because the weakness of questionnaire-based tools lies within the scoring and implications components, we recommend attending to how assessors are selected and ensuring these assessors' adequate exposure to the physician in question when using questionnaire-based tools.

The utility of each assessment method is always a compromise between various aspects of quality, such as validity evidence.[142] Hence, combining questionnaire-based tools with other assessment methods that have sufficient evidence for other components of the validity argument provides a more meaningful assessment program compared with using any single method in isolation from another. We cannot make general recommendations on which tool to use. Identifying one single best tool proved to be challenging because of the context- and specialty-specific character of the reviewed tools. Potential users of questionnaire-based tools should select the tool that best serves their intended assessment purpose, based on the available validity evidence and the value ascribed to that evidence. The complete overview of validity evidence per tool (Supplemental Digital Appendix 2, available at http://links.lww.com/ACADMED/A677) may serve as a guide to facilitate the selection process.

To understand and discern which tools are needed in a full physician assessment program, examination of the content of questionnaire-based tools in relation to their constructive alignment is needed; for example, what is the tool's relationship to competency frameworks? Exploring a more programmatic or comprehensive and holistic approach to assessing physicians' clinical and teaching performance may be worthwhile. A *meaningful* assessment of physicians requires a combination of various tools; all tools need not be perfect, but the combination of tools should be thoughtful.[138]

Limitations and strengths

This study has some limitations. First, we may not have identified all studies, and therefore our review may be incomplete and potentially biased. Second, only one author (M.W.vdM.) reviewed the initial abstracts in the first screening stage of the process. Third, by considering only the weakest assumptions stated a priori, we might have taken a somewhat deductive approach to collecting the validity evidence for the questionnaire-based tools. Given all the validity frameworks, we could have selected multiple ways to seek validity evidence; we made pragmatic choices to avoid a never-ending process wherein we would have interpreted and incorporated every piece of validity evidence available and then continually calculated a new score.[143] There is considerable heterogeneity in the identified studies in terms of study design, quality, and context, which made the assimilation of evidence challenging, yet not impossible due to the argument-based approach to validity that we used. Using our argument-based approach, we were able to collect and assimilate different types of evidence—from quantitative, as well as qualitative, studies.[142,144] As far as we are aware, this is the first review to rigorously examine questionnaire-based tools with an argument-based approach to validity. We tackled the central issue in the validity debate, giving more weight to the scoring and implications components of the argument than to the extrapolation and generalization components, because the former are especially needed for summative uses of these types of tools. Given the argument-based approach we used, which evaluates the argument for validity by weighing the components differently and prioritizing evidence based on the intended use of the tool,[13,16] we have provided a state-of-the-art perspective of validity.

Conclusions

For several years, society has increasingly focused on the assessment of physicians' professional performance to support physicians in delivering optimal patient care, training competent future doctors, and conducting innovative research. Questionnaire-based tools have played an important role in meeting this professional and public need, yet the validity evidence for these tools has some flaws. Some of these flaws are inherent to questionnaire-based tools, and some tools are poorly designed, thus providing insufficient evidence to support their use. We therefore feel that the way forward is twofold: (1) to continue the collection of evidence to support the validity argument of existing tools, and (2) to explore which combination of questionnaire-based tools can collectively contribute to a valid and meaningful assessment of physicians' performance. This dual approach may be instrumental in building an effective

toolbox to help develop a workforce of high-performing physicians who educate the next generation of physicians, conduct research, and deliver high-quality health care.

Acknowledgments: The authors wish to thank the clinical librarian Faridi van Etten-Jamaludin from the Academic Medical Center University of Amsterdam, Amsterdam, The Netherlands, for her help in setting up a thorough search strategy.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

Previous presentations: This review was presented at the 2018 biannual OTTAWA-International Conference on Medical Education Conference in Abu Dhabi, United Arab Emirates, March 14, 2018; at the 2018 annual Association of Medical Educators of Europe Conference in Basel, Switzerland on August 28, 2018 as an oral presentation; and at the 2018 annual Nederlandse Vereniging voor Medisch Onderwijs Congress in Egmond aan Zee, The Netherlands, November 16, 2018.

M.W. van der Meulen is PhD candidate, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands, and member, Professional Performance Research Group, Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; ORCID: https://orcid.org/0000-0003-3636-5469.

A. Smirnova is PhD graduate and researcher, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands, and member, Professional Performance Research Group, Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; ORCID: https://orcid.org/0000-0003-4491-3007.

S. Heeneman is professor, Department of Pathology, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands; ORCID: https://orcid.org/0000-0002-6103-8075.

M.G.A. oude Egbrink is professor, Department of Physiology, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands; ORCID: https://orcid.org/0000-0002-5530-6598.

C.P.M. van der Vleuten is professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, The Netherlands; ORCID: https://orcid.org/0000-0001-6802-3119.

K.M.J.M.H. Lombarts is professor, Professional Performance Research Group, Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; ORCID: https://orcid.org/0000-0001-6167-0620.

References

1 Daouk-Öyry L, Zaatari G, Sahakian T, Rahal Alameh B, Mansour N. Developing a competency framework for academic physicians. Med Teach. 2017;39:269–277.
2 Milner RJ, Gusic ME, Thorndyke LE. Perspective: Toward a competency framework for faculty. Acad Med. 2011;86:1204–1210.
3 Mackillop LH, Crossley J, Vivekananda-Schmidt P, Wade W, Armitage M. A single generic multi-source feedback tool for revalidation of all UK career-grade doctors: Does one size fit all? Med Teach. 2011;33:e75–e83.
4 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. BMJ. 2004;328:1240.
5 Ramsey PG, Wenrich MD. Peer ratings. An assessment tool whose time has come. J Gen Intern Med. 1999;14:581–582.
6 Al Ansari A, Donnon T, Al Khalifa K, Darwish A, Violato C. The construct and criterion validity of the multi-source feedback process to assess physician performance: A meta-analysis. Adv Med Educ Pract. 2014;5:39–51.
7 Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med. 2005;20:1159–1164.
8 Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med. 2004;19:971–977.
9 Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. Acad Med. 2014;89:511–516.
10 Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the quality of clinical teachers: A systematic review of content and quality of questionnaires for assessing clinical teachers. J Gen Intern Med. 2010;25:1337–1345.
11 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–575.
12 Stevens S, Read J, Baines R, Chatterjee A, Archer J. Validation of multisource feedback in assessing medical performance: A systematic review. J Contin Educ Health Prof. 2018;38:262–268.
13 Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50:1–73.
14 Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-Clinical Evaluation Exercise: A review of the research. Acad Med. 2010;85:1453–1461.
15 Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ. 2009;339:b2535.
16 Kane MT. An argument-based approach to validity. Psychol Bull. 1992;112:527–535.
17 Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. Adv Health Sci Educ Theory Pract. 2012;17:165–181.
18 American Education Research Association; American Psychological Association; National Council on Measurement in Education; The Joint Committee on Standards for Education and Psychological Testing. Standards for Educational and Psychological Testing. Washington, DC: American Education Research Association; 1999.
19 Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. Eval Health Prof. 1989;12:409–423.
20 Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Ann Intern Med. 1989;110:719–726.
21 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655–1660.
22 Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. Acad Med. 1993;68:680–687.
23 Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med. 1996;71:364–370.
24 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med. 1997;72(10 suppl 1):S82–S84.
25 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: The effect of individual feedback. Acad Med. 1999;74:702–714.
26 Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: The physician achievement review. CMAJ. 1999;161:52–57.
27 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med. 2002;77(10 suppl):S64–S66.
28 Lockyer J, Violato C, Fidler H. Likelihood of change: A study assessing surgeon use of multisource feedback data. Teach Learn Med. 2003;15:168–174.
29 Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: A pilot study. Acad Med. 2003;78(10 suppl):S42–S44.
30 Violato C, Lockyer J, Fidler H. Multisource feedback: A method of assessing surgical practice. BMJ. 2003;326:546–548.
31 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med. 2004;79(10 suppl):S5–S8.
32 Elwyn G, Lewis M, Evans R, Hutchings H. Using a "peer assessment questionnaire" in primary medical care. Br J Gen Pract. 2005;55:690–695.
33 Rosenbaum ME, Ferguson KJ, Kreiter CD, Johnson CA. Using a peer evaluation system to assess faculty performance and competence. Fam Med. 2005;37:429–433.
34 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. Med Educ. 2005;39:497–504.

35 Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Can J Anaesth. 2006;53:33–39.
36 Lockyer JM, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. Acad Emerg Med. 2006;13:1296–1303.
37 Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. Pediatrics. 2006;117:796–802.
38 Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Challenges in multisource feedback: Intended and unintended outcomes. Med Educ. 2007;41:583–591.
39 Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: An evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care. 2008;17:187–193.
40 Crossley J, McDonnell J, Cooper C, McAvoy P, Archer J, Davies H. Can a district hospital assess its doctors for re-licensure? Med Educ. 2008;42:359–363.
41 Lelliott P, Williams R, Mears A, et al. Questionnaires for 360-degree assessment of consultant psychiatrists: Development and psychometric properties. Br J Psychiatry. 2008;193:156–160.
42 Lockyer JM, Violato C, Fidler HM. Assessment of radiology physicians by a regulatory authority. Radiology. 2008;247:771–778.
43 Sargeant J, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract. 2008;13:275–288.
44 Violato C, Lockyer JM, Fidler H. Changes in performance: A 5-year longitudinal study of participants in a multi-source feedback programme. Med Educ. 2008;42:1007–1013.
45 Violato C, Lockyer JM, Fidler H. Assessment of psychiatrists in practice through multisource feedback. Can J Psychiatry. 2008;53:525–533.
46 Hess BJ, Lynn LA, Holmboe ES, Lipner RS. Toward better care coordination through improved communication with referring physicians. Acad Med. 2009;84(10 suppl):S109–S112.
47 Lockyer JM, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. Arch Pathol Lab Med. 2009;133:1301–1308.
48 Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ. 2009;43:757–766.
49 Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: A link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract. 2009;14:399–410.
50 Campbell J, Narayanan A, Burford B, Greco M. Validation of a multi-source feedback tool for use in general practice. Educ Prim Care. 2010;21:165–179.
51 Shepherd A, Lough M. What is a good general practitioner (GP)? The development and evaluation of a multi-source feedback instrument for GP appraisal. Educ Prim Care. 2010;21:149–164.
52 Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. Med Educ. 2011;45:886–893.
53 Campbell JL, Roberts M, Wright C, et al. Factors associated with variability in the assessment of UK doctors' professionalism: Analysis of survey results. BMJ. 2011;343:d6212.
54 Mackillop L, Parker-Swift J, Crossley J. Getting the questions right: Non-compound questions are more reliable than compound questions on matched multi-source feedback instruments. Med Educ. 2011;45:843–848.
55 Sargeant J, Macleod T, Sinclair D, Power M. How do physicians assess their family physician colleagues' performance? Creating a rubric to inform assessment and feedback. J Contin Educ Health Prof. 2011;31:87–94.
56 Bhogal HK, Howell E, Torok H, Knight AM, Howell E, Wright S. Peer assessment of professional performance by hospitalist physicians. South Med J. 2012;105:254–258.
57 Hill JJ, Asprey A, Richards SH, Campbell JL. Multisource feedback questionnaires in appraisal and for revalidation: A qualitative study in UK general practice. Br J Gen Pract. 2012;62:e314–e321.
58 Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. BMC Health Serv Res. 2012;12:80.
59 Overeem K, Wollersheimh HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Factors predicting doctors' reporting of performance change in response to multisource feedback. BMC Med Educ. 2012;12:52.
60 Wright C, Richards SH, Hill JJ, et al. Multisource feedback in evaluating the performance of doctors: The example of the UK General Medical Council patient and colleague questionnaires. Acad Med. 2012;87:1668–1678.
61 Vinod SK, Lonergan DM. Multisource feedback for radiation oncologists. J Med Imaging Radiat Oncol. 2013;57:384–389.
62 Warner DO, Sun H, Harman AE, Culley DJ. Feasibility of patient and peer surveys for maintenance of certification among diplomates of the American Board of Anesthesiology. J Clin Anesth. 2015;27:290–295.
63 Al Ansari A, Al Meer A, Althawadi M, Henari D, Al Khalifa K. Cross-cultural challenges in assessing medical professionalism among emergency physicians in a Middle Eastern Country (Bahrain): Feasibility and psychometric properties of multisource feedback. Int J Emerg Med. 2016;9:2.
64 Metz R, Haring O. An apparent relationship between the seniority of faculty members and their ratings as bedside teachers. J Med Educ. 1966;41:1057–1062.
65 Tiberius RG, Sackin HD, Slingerland JM, Jubas K, Bell M, Matlow A. The influence of student evaluative feedback on the improvement of clinical teaching. J High Educ. 1989;60:665–681.
66 McLeod P. Faculty perspectives of a valid and reliable clinical tutor evaluation program. Eval Health Prof. 1991;14:333–342.
67 Tortolani AJ, Risucci DA, Rosati RJ. Resident evaluation of surgical faculty. J Surg Res. 1991;51:186–191.
68 Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ. Reliability and accuracy of resident evaluations of surgical faculty. Eval Health Prof. 1992;15:313–324.
69 Ramsbottom-Lucier MT, Gillmore GM, Irby DM, Ramsey PG. Evaluation of clinical teaching by general internal medicine faculty in outpatient and inpatient settings. Acad Med. 1994;69:152–154.
70 Schum TR, Yindra KJ. Relationship between systematic feedback to faculty and ratings of clinical teaching. Acad Med. 1996;71:1100–1102.
71 Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ. The reliability of medical student ratings of clinical teaching. Eval Health Prof. 1997;20:343–352.
72 Litzelman DK, Westmoreland GR, Skeff KM, Stratos GA. Factorial validation of an educational framework using residents' evaluations of clinician–educators. Acad Med. 1999;74(10 suppl):S25–S27.
73 Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. Acad Med. 2000;75:161–166.
74 Steiner IP, Franc-Law J, Kelly KD, Rowe BH. Faculty evaluation by residents in an emergency medicine program: A new evaluation instrument. Acad Emerg Med. 2000;7:1015–1021.
75 Shea JA, Bellini LM. Evaluations of clinical faculty: The impact of level of learner and time of year. Teach Learn Med. 2002;14:87–91.
76 de Groot J, Brunet A, Kaplan AS, Bagby M. A comparison of evaluations of male and female psychiatry supervisors. Acad Psychiatry. 2003;27:39–43.
77 Donner-Banzhoff N, Merle H, Baum E, Basler HD. Feedback for general practice trainers: Developing and testing a standardised instrument using the importance-quality-score method. Med Educ. 2003;37:772–777.
78 Steiner IP, Yoon PW, Kelly KD, et al. Resident evaluation of clinical teachers based on teachers' certification. Acad Emerg Med. 2003;10:731–737.
79 Kripalani S, Pope AC, Rask K, et al. Hospitalists as teachers. J Gen Intern Med. 2004;19:8–15.
80 Maker VK, Curtis KD, Donnelly MB. Faculty evaluations: Diagnostic and therapeutic. Curr Surg. 2004;61:597–601.
81 Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians: A new instrument for today's teaching hospitals. J Gen Intern Med. 2004;19:766–771.
82 Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. Fam Med. 2005;37:43–47.

83 Beckman TJ, Mandrekar JN. The interpersonal, cognitive and efficiency domains of clinical teaching: Construct validity of a multi-dimensional scale. Med Educ. 2005;39:1221–1229.

84 Steiner IP, Yoon PW, Kelly KD, et al. The influence of residents training level on their evaluation of clinical teaching faculty. Teach Learn Med. 2005;17:42–48.

85 Beckman TJ, Cook DA, Mandrekar JN. Factor instability of clinical teaching assessment scores among general internists and cardiologists. Med Educ. 2006;40:1209–1216.

86 Maker VK, Lewis MJ, Donnelly MB. Ongoing faculty evaluations: Developmental gain or just more pain? Curr Surg. 2006;63:80–84.

87 Mourad O, Redelmeier DA. Clinical teaching and clinical outcomes: Teaching capability and its association with patient outcomes. Med Educ. 2006;40:637–644.

88 Silber C, Novielli K, Paskin D, et al. Use of critical incidents to develop a rating form for resident evaluation of faculty teaching. Med Educ. 2006;40:1201–1208.

89 Bierer SB, Hull AL. Examination of a clinical teaching effectiveness instrument used for summative faculty assessment. Eval Health Prof. 2007;30:339–361.

90 Kelly SP, Shapiro N, Woodruff M, Corrigan K, Sanchez LD, Wolfe RE. The effects of clinical workload on teaching in the emergency department. Acad Emerg Med. 2007;14:526–531.

91 McOwen KS, Bellini LM, Guerra CE, Shea JA. Evaluation of clinical faculty: Gender and minority implications. Acad Med. 2007;82(10 suppl):S94–S96.

92 McOwen KS, Bellini LM, Shea JA. Residents' ratings of clinical excellence and teaching effectiveness: Is there a relationship? Teach Learn Med. 2007;19:372–377.

93 Zuberi RW, Bordage G, Norman GR. Validation of the SETOC instrument—Student evaluation of teaching in outpatient clinics. Adv Health Sci Educ Theory Pract. 2007;12:55–69.

94 de Oliveira Filho GR, Dal Mago AJ, Garcia JH, Goldschmidt R. An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. Anesth Analg. 2008;107:1316–1322.

95 Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The development of an instrument for evaluating clinical teachers: Involving stakeholders to determine content validity. Med Teach. 2008;30:e272–e277.

96 Lombarts KM, Bucx MJ, Arah OA. Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. Anesthesiology. 2009;111:709–716.

97 Shea JA, Bellini LM, McOwen KS, Norcini JJ. Setting standards for teaching evaluation data: An application of the contrasting groups method. Teach Learn Med. 2009;21:82–86.

98 Baker K. Clinical teaching improves with resident evaluation and feedback. Anesthesiology. 2010;113:693–703.

99 Beckman TJ, Reed DA, Shanafelt TD, West CP. Impact of resident well-being and empathy on assessments of faculty physicians. J Gen Intern Med. 2010;25:52–56.

100 Colletti JE, Flottemesch TJ, O'Connell TA, Ankel FK, Asplin BR. Developing a standardized faculty evaluation in an emergency medicine residency. J Emerg Med. 2010;39:662–668.

101 Lombarts KM, Heineman MJ, Arah OA. Good clinical teachers likely to be specialist role models: Results from a multicenter cross-sectional survey. PLoS One. 2010;5:e15202.

102 Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med. 2010;85:1732–1738.

103 Arah OA, Hoekstra JB, Bos AP, Lombarts KM. New tools for systematic evaluation of teaching qualities of medical faculty: Results of an ongoing multi-center survey. PLoS One. 2011;6:e25983.

104 Logio LS, Monahan P, Stump TE, Branch WT Jr, Frankel RM, Inui TS. Exploring the psychometric properties of the humanistic teaching practices effectiveness questionnaire, an instrument to measure the humanistic qualities of medical teachers. Acad Med. 2011;86:1019–1025.

105 Nation JG, Carmichael E, Fidler H, Violato C. The development of an instrument to assess clinical teaching with linkage to CanMEDS roles: A psychometric analysis. Med Teach. 2011;33:e290–e296.

106 van der Leeuw R, Lombarts K, Heineman MJ, Arah O. Systematic evaluation of the teaching qualities of obstetrics and gynecology faculty: Reliability and validity of the SETQ tools. PLoS One. 2011;6:e19142.

107 Zibrowski EM, Myers K, Norman G, Goldszmidt MA. Relying on others' reliability: Challenges in clinical teaching assessment. Teach Learn Med. 2011;23:21–27.

108 Arah OA, Heineman MJ, Lombarts KM. Factors influencing residents' evaluations of clinical faculty member teaching qualities and role model status. Med Educ. 2012;46:381–389.

109 Boerebach BC, Arah OA, Busch OR, Lombarts KM. Reliable and valid tools for measuring surgeons' teaching performance: Residents' vs. self evaluation. J Surg Educ. 2012;69:511–520.

110 Egbe M, Baker P. Development of a multisource feedback instrument for clinical supervisors in postgraduate medical training. Clin Med (Lond). 2012;12:239–243.

111 Fluit C, Bolhuis S, Grol R, et al. Evaluation and feedback for effective clinical teaching in postgraduate medical education: Validation of an assessment instrument incorporating the CanMEDS roles. Med Teach. 2012;34:893–901.

112 Schönrock-Adema J, Boendermaker PM, Remmelts P. Opportunities for the CTEI: Disentangling frequency and quality in evaluating teaching behaviours. Perspect Med Educ. 2012;1:172–179.

113 Archer J, Swanwick T, Smith D, O'Keeffe C, Cater N. Developing a multisource feedback tool for postgraduate medical educational supervisors. Med Teach. 2013;35:145–154.

114 Backeris ME, Patel RM, Metro DG, Sakai T. Impact of a productivity-based compensation system on faculty clinical teaching scores, as evaluated by anesthesiology residents. J Clin Anesth. 2013;25:209–213.

115 Fluit CR, Feskens R, Bolhuis S, Grol R, Wensing M, Laan R. Repeated evaluations of the quality of clinical teaching by residents. Perspect Med Educ. 2013;2:87–94.

116 Hindman BJ, Dexter F, Kreiter CD, Wachtel RE. Determinants, associations, and psychometric properties of resident assessments of anesthesiologist operating room supervision. Anesth Analg. 2013;116:1342–1351.

117 van der Leeuw RM, Overeem K, Arah OA, Heineman MJ, Lombarts KM. Frequency and determinants of residents' narrative feedback on the teaching performance of faculty: Narratives in numbers. Acad Med. 2013;88:1324–1331.

118 van der Leeuw RM, Slootweg IA, Heineman MJ, Lombarts KM. Explaining how faculty members act upon residents' feedback to improve their teaching performance. Med Educ. 2013;47:1089–1098.

119 Kikukawa M, Stalmeijer RE, Emura S, Roff S, Scherpbier AJ. An instrument for evaluating clinical teaching in Japan: Content validity and cultural sensitivity. BMC Med Educ. 2014;14:179.

120 Lases SS, Arah OA, Pierik EG, Heineman E, Lombarts MJ. Residents' engagement and empathy associated with their perception of faculty's teaching performance. World J Surg. 2014;38:2753–2760.

121 Lee SM, Lee MC, Reed DA, et al. Success of a faculty development program for teachers at the Mayo Clinic. J Grad Med Educ. 2014;6:704–708.

122 Lombarts KM, Heineman MJ, Scherpbier AJ, Arah OA. Effect of the learning climate of residency programs on faculty's teaching performance as evaluated by residents. PLoS One. 2014;9:e86512.

123 Scheepers RA, Lombarts KM, van Aken MA, Heineman MJ, Arah OA. Personality traits affect teaching performance of attending physicians: Results of a multi-center observational study. PLoS One. 2014;9:e98107.

124 Young ME, Cruess SR, Cruess RL, Steinert Y. The Professionalism Assessment of Clinical Teachers (PACT): The reliability and validity of a novel tool to evaluate professional and clinical teaching behaviors. Adv Health Sci Educ Theory Pract. 2014;19:99–113.

125 Da Dalt L, Anselmi P, Furlan S, et al. Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. Ital J Pediatr. 2015;41:2.

126 Mintz M, Southern DA, Ghali WA, Ma IW. Validation of the 25-Item Stanford Faculty Development Program Tool on Clinical Teaching Effectiveness. Teach Learn Med. 2015;27:174–181.

127 Robinson RL. Hospitalist workload influences faculty evaluations by internal medicine clerkship students. Adv Med Educ Pract. 2015;6:93–98.

128 Boerebach BC, Lombarts KM, Arah OA. Confirmatory factor analysis of the System for Evaluation of Teaching Qualities (SETQ) in graduate medical training. Eval Health Prof. 2016;39:21–32.

129 Dexter F, Szeluga D, Masursky D, Hindman BJ. Written comments made by anesthesia

residents when providing below average scores for the supervision provided by the faculty anesthesiologist. Anesth Analg. 2016;122:2000–2006.
130 Huete Á, Julio R, Rojas V, et al. Evaluation of radiology teachers' performance and identification of the "best teachers" in a residency program: Mixed methodology and pilot study of the MEDUC-RX32 questionnaire. Acad Radiol. 2016;23:779–788.
131 Lombarts KM, Ferguson A, Hollmann MW, Malling B, Arah OA; SMART Collaborators. Redesign of the System for Evaluation of Teaching Qualities in Anesthesiology Residency Training (SETQ Smart). Anesthesiology. 2016;125:1056–1065.
132 Scheepers RA, Arah OA, Heineman MJ, Lombarts KM. How personality traits affect clinician–supervisors' work engagement and subsequently their teaching performance in residency training. Med Teach. 2016;38:1105–1111.
133 Van Der Leeuw RM, Boerebach BC, Lombarts KM, Heineman MJ, Arah OA. Clinical teaching performance improvement of faculty in residency training: A prospective cohort study. Med Teach. 2016;38:464–470.
134 Wingo MT, Halvorsen AJ, Beckman TJ, Johnson MG, Reed DA. Associations between attending physician workload, teaching effectiveness, and patient safety. J Hosp Med. 2016;11:169–173.
135 van der Hem-Stokroos HH, van der Vleuten CP, Daelmans HE, Haarman HJ, Scherpbier AJ. Reliability of the clinical teaching effectiveness instrument. Med Educ. 2005;39:904–910.
136 Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics: A Primer. New York, NY: John Wiley & Sons; 2016.
137 Kane MT. Current concerns in validity theory. J Educ Meas. 2001;38:319–342.
138 van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34:205–214.
139 Goldstein MJ, Lunn MR, Peng L. What makes a top research medical school? A call for a new model to evaluate academic physicians and medical school performance. Acad Med. 2015;90:603–608.
140 Patel VM, Ashrafian H, Bornmann L, et al. Enhancing the h index for the objective assessment of healthcare researcher performance and impact. J R Soc Med. 2013;106:19–29.
141 Federatie Medisch Specialisten. Position Paper: The Medical Specialist as a Scientist [in Dutch]. Utrecht, The Netherlands: Royal Dutch Medical Association; December 2017. https://www.demedischspecialist.nl/sites/default/files/position%20paper%20De%20medisch%20specialist%20als%20wetenschapper.pdf. Accessed April 4, 2019.
142 Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. Med Educ. 2012;46:38–48.
143 St-Onge C, Young M, Eva KW, Hodges B. Validity: One word with a plurality of meanings. Adv Health Sci Educ Theory Pract. 2017;22:853–867.
144 Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. Acad Med. 2016;91:1359–1369.