COMPREHENSIVE REVIEW

On the Assessment of Paramedic Competence: A Narrative Review with Practice Implications

W. Tavares, PhD;[1-5] S. Boet, MD, PhD[6]

1. Centennial College, Paramedic Program, Toronto, Ontario, Canada
2. McMaster University, Faculty of Medicine, Division of Emergency Medicine, Hamilton, Ontario, Canada
3. ORNGE Transport Medicine, Mississauga, Ontario, Canada
4. York Region Paramedic Services, Newmarket, Ontario, Canada
5. Paramedic Association of Canada, Ottawa, Ontario, Canada
6. University of Ottawa, Skills and Simulation Centre and Department of Anesthesiology, Ottawa Hospital, Ottawa, Ontario, Canada

Correspondence:
Walter Tavares, PhD
Centennial College
P.O. Box 631, Station A
Toronto, Ontario, Canada M1K 5E9
E-mail: wtavares@centennialcollege.ca

Conflicts of interest: The authors report no conflicts of interest.

Keywords: clinical competence; educational measurement; Emergency Medical Services; Emergency Medical Technicians

Abbreviations:
ECG: electrocardiogram
GRS: global rating scale
OSCE: objective structured clinical examination
PBA: performance-based assessment
RBA: rater-based assessment
SBA: simulation-based assessment
WBA: workplace-based assessment

Received: May 2, 2015
Revised: June 23, 2015
Accepted: July 25, 2015

Online publication: November 30, 2015

doi:10.1017/S1049023X15005166

Abstract

Introduction: Paramedicine is experiencing significant growth in scope of practice, autonomy, and role in the health care system. Despite clinical governance models, the degree to which paramedicine ultimately can be safe and effective will be dependent on the individuals the profession deems suited to practice. This creates an imperative for those responsible for these decisions to ensure that assessments of paramedic competence are indeed accurate, trustworthy, and defensible.

Purpose: The purpose of this study was to explore and synthesize relevant theoretical foundations and literature informing best practices in performance-based assessment (PBA) of competence, as it might be applied to paramedicine, for design or evaluation of assessment programs.

Methods: A narrative review methodology was applied to focus intentionally, but broadly, on purpose relevant, theoretically derived research that could inform assessment protocols in paramedicine. Primary and secondary studies from a number of health professions that contributed to and informed best practices related to the assessment of paramedic clinical competence were included and synthesized.

Results: Multiple conceptual frameworks, psychometric requirements, and emerging lines of research are forwarded. Seventeen practice implications are derived to promote understanding as well as best practices and evaluation criteria for educators, employers, and/or licensing/certifying bodies when considering the assessment of paramedic competence.

Conclusions: The assessment of paramedic competence is a complex process requiring an understanding, appreciation for, and integration of conceptual and psychometric principles. The field of PBA is advancing rapidly with numerous opportunities for research.

Tavares W, Boet S. On the assessment of paramedic competence: a narrative review with practice implications. *Prehosp Disaster Med*. 2016;31(1):64-73.

Introduction

Paramedicine has experienced significant growth in recent years, extending traditional emergency or acute care roles with increases in breadth of practice and autonomy. At the same time, researchers have warned of the increasing risk to patient safety that this creates, especially as it relates to demands on clinical reasoning and decision making, but also where additional technical and other non-technical skills are required.[1-3] Safe paramedic practice has largely been dependent on clinical governance models (eg, medical oversight and/or the use of medical directives or clinical guidelines). However, these likely are limited in their ability to account for the degree of medical ambiguity, limited diagnostics, diverse contexts and circumstances, and growing patient groups paramedics increasingly are challenged with.[4] Therefore, the extent to which paramedicine ultimately can be safe will be dependent not only on systems (eg, infrastructure, oversight, equipment, and process), but also on the safe clinical practice of those individuals serving within the profession.[1,3] As such, making accurate and trustworthy decisions regarding paramedic clinical proficiency, especially at the entry to practice level, can have significant implications for patient safety overall. While no one assessment method can be responsible for informing all aspects of clinical practice or competence,[5] in many instances, performance-based assessments (PBAs; ie, where candidates exhibit behaviors in response to clinical challenges) have and continue to serve an integral role.[6,7] Therefore, the goal of this review was to provide the paramedic community with literature-based foundations by which to establish optimal PBA strategies.

Goals associated with PBAs can include: (a) differentiating between levels of performance; (b) making accurate determinations regarding the achievement of predefined competencies; (c) detecting the ability to adaptively apply those competencies; and/or (d) to make accurate predictions regarding future clinical performance in novel contexts. However, a review of the paramedicine literature reveals a paucity of research aimed at providing evidence for, or informing, best practices related to achieving these goals.[8] What research does exists has been limited to identifying performance deficiencies in specific patient types (eg, pediatric resuscitation),[9,10] as part of correlational studies involving a single PBA,[11] studies assessing isolated technical skills[12-19] or non-technical skills,[20] or as part of outcome measures in research studying various interventions (eg, comparing intubation interventions, stress, and so on).[11,21-28] Furthermore, two systematic reviews in health professions education, the first identifying tools used for assessment of clinical competence and the other evaluating the use of simulation for the purpose of assessment, both failed to reveal any meaningful evidence related to paramedicine.[29,30] Other recent research has been limited to scale development and validation, providing little to overall process and best practices.[31-34]

Despite the lack of profession-specific evidence, the broader health professions and assessment literature is extensive. For instance, a number of reviews have been conducted previously summarizing evidence and meaningful perspectives in the assessment of clinical competence.[35-50] This expansive body of literature can inform and contribute process, but context remains an important feature. Many assessment principles, including reliability, validity, and the role of "subjective" rater-based assessments (RBAs) may be directly applicable, yet are seldom applied in paramedicine. Therefore, this narrative review aimed to provide a synthesis of the relevant theoretical foundations and literature informing PBA as it might be applied to paramedicine. As such, this review was not based on a single, highly specific, or standardized search, nor was it an exhaustive review of the PBA literature. Instead, as per the narrative tradition, this study focused intentionally on various, but consistent, theoretically derived research, as well as paramedic-specific research (where applicable) that collectively can inform assessment best practices in paramedicine. Reviews of this type are well suited for comprehensive topic areas that require a holistic interpretation and integration of multiple and diverse existing theories and pluralities of scholarly work that are not amenable to quantitative analyses or results or highly specific search criteria. The authors of this report tried to be comprehensive in their review and limited bias by remaining broad in their inclusion of the literature and by providing readers with the information necessary to form their own interpretations. To provide boundaries in this search, this study focused mainly on summative assessments where the adaptive integration of multiple competencies for the purpose of making entry to practice decisions and optimizing predictions of future clinical performance in novel contexts is desired. The focus was on current understanding of various theoretical frameworks that support recommendations educators, employers, and/or licensing or certifying bodies may use (without being overly prescriptive) to evaluate or design their assessment processes, while also exploring emerging lines of inquiry in PBA literature.

Report and Discussion

The Act of Rater-based Assessments of Competence

Before exploring features associated with optimal PBA, it is helpful to explore the act of RBA of clinical competence. In its simplest form, RBA begins with presenting candidates with a clinical challenge. Depending on the context, the clinical interaction may be created deliberately, controlled, or selected (eg, simulation-based settings) or not (eg, field settings). Next, the candidate interacts with the clinical challenge, selecting and adaptively applying and demonstrating various competencies as the case warrants, exhibited as words, actions (or inactions), events, or interactions. Candidates do this fluidly in response to changing parameters (eg, changes in patient condition and implicit feedback). A rater must then consider the reams of information, process and interpret all of this in reference to some standard, then ultimately translate this information into a rating, categorical judgment, and/or some form of narrative. The degree to which assessment goals (described above) are achieved, or are even feasible, will depend on how these individual features are understood and applied.

Understanding Competence

In order to inform assessment of competence, it is helpful to understand the way in which competence has been conceptualized and discussed. Kane defined competence as "the degree to which an individual can use the knowledge, skills, and judgment (some have since added attitudes) associated with the profession to perform effectively in a domain of possible encounters defining the scope of professional practice."[51] Other definitions similarly capture the individual's features, while drawing in the breadth and boundaries of the profession.[52] This definition suggests that understanding the degree to which a paramedic is competent requires that candidates demonstrate (ie, "use" or "perform") various competencies adaptively across a number of contexts and patient types. Further, that knowledge, skill, and judgment not necessarily be limited to any one aspect of practice, rather that assessment designers optimally represent the profession in the assessment effort. Kane makes no distinction over technical or non-technical aspects of practice, and therefore, both may need to be included as they are important features of paramedic practice.[53,54]

In assessment terms, competence is related closely to the concept of a universe score (also referred to as a "true" score). A universe score refers to the long-run hypothetical mean a candidate would receive, absent of measurement error, across all possible observations.[55,56] In other words, if a candidate could be observed across all possible patient encounters representative of the field of paramedicine, one would essentially understand, or in effect, know that candidate's "true" performance ability or degree of competence. However, this carries obvious feasibility challenges, and measurement error (ie, anything that results in deviations from this true score) is always present to some extent. In this universe score framework (which adopts a positivist perspective), "true" performance ability exists within every candidate to varying degrees, and it can be measured. However, "performance ability" is an abstract construct (like intelligence or motivation). Because of this, performance ability can only ever be inferred based on observation of behaviors in response to some form of stimuli. Using Kane's views on competence and the concept of a universe score, what follows is a review of a number of conceptual frameworks with practice implications to consider when making decisions regarding assessment design. It is worth noting that the assessment community (and thus, this review) has been focused mainly, for over 50 years, on psychometric principles when thinking about, and on assessments practices in, health professions

https://doi.org/10.1017/S1049023X15005166 Published online by Cambridge University Press

education. However, researchers have begun to explore a post-psychometric era in assessment, which is discussed below under the heading "Emerging Areas of Assessment Research."

Millers Pyramid of Competence

In describing how best to conceptualize the assessment of competence, Miller proposed an assessment framework that consists of four levels of performance (often illustrated as a pyramid), each representing an increasingly complex stimulus and response format.[6,36] At the base of the pyramid is Level 1, referred to as "knows," followed by Level 2, or "knows how." These levels involve assessment of cognition, mainly in the form of declarative and applied knowledge, respectively. These can be measured efficiently and effectively using stimuli and response formats (eg, texted-based questions with multiple-choice options) that do not require any complex performance or behaviors. For example, describing coronary blood flow and electrocardiogram (ECG) patterns related to various areas of the heart would be at the level of "knows." Selecting care plans based on ECG findings and/or associated manifestations would be at the level of "knows how." Level 3 and Level 4, referred to as "shows how" and "does," respectively, require individuals to demonstrate behaviors in response to clinical challenges either in simulation ("shows how") or with real patients in real clinical contexts ("does").

Practice Implications—First, no one-assessment strategy can capture all levels of Miller's Pyramid. This suggests the need for programmatic assessment with targeted strategies depending on areas of focus.[36] Assess at higher levels of Miller's pyramid what cannot be assessed more efficiently at lower levels.

Second, an underlying principle is that while performance at one level (eg, Level 2 "knows how") may be dependent on the level below it (ie, Level 1 "knows"), that same level does not necessarily predict performance at the level above (ie, at Level 3 "shows how").[6,57]

Third, competence involves a progression of knowledge, skills, and abilities, best measured in settings where the assessment context closely parallels the environment in which future clinical performance is expected to occur, mainly to minimize leaps in extrapolation in an inference-based model.[40] Ideally, all clinical competence decisions ultimately would include assessment of behaviors exhibited in real clinical contexts with real patients (ie, at the "does" level).[57,58] However, in paramedic contexts in particular, work-based assessments (WBAs) are often associated with a number of challenges, such as lack of control over content (potentially leading to inappropriate case variability and/or complexity or a situation in which the sample of cases used for summative decisions may represent inadequately the clinical domain), interruptions due to patient safety concerns, and undue influence by many uncontrollable contextual factors.[59]

As result of challenges associated with WBA, many have advocated for, and adopted, simulation-based assessments (SBAs;[59] ie, the "shows how" level) where ecological validity (ie, similarity between assessment and work-based settings), standardization, elimination of patient safety barriers, and complete control over content can be exercised. However, SBA, only ever being a surrogate of reality, requires evidence of both reliability and validity.

Reliability

Ensuring a degree of reliability promotes trustworthiness in the PBA process. Reliability refers to the ability to differentiate consistently between candidates with both consistency and differentiation being important.[60,61] Consistency suggests that scores on a performance assessment would be relatively similar between raters (inter-rater reliability), within raters (intra-rater reliability), if assessed repeatedly (test-retest reliability), and/or among stations (inter-station reliability). Differentiation suggests the assessment process is designed and implemented such that differences between candidates (if present) can be detected. Reliability is also an indicator of the amount of error associated with the process.[61,62] It is calculated by taking the ratio of subject variance (eg, differences between candidates) to subject variance plus error, with a score of "1" indicating perfect reliability.[63] In an ideal sense, there would never be any measurement error in the assessment process and the reliability would always equal "1." However, in reality, this is never the case. Therefore, a common strategy is to identify sources of measurement error and apply whatever strategies are available to mitigate them. Examples are provided below. Poor reliability suggests the risk of making classification errors is high. This renders the assessment process useless since incorrect decisions regarding a paramedic candidate can ultimately be made as a result, with significant dangerous downstream implications for the patients/public and profession.

As suggested above, two main sources of threat to reliability include factors contributing to poor differentiation and/or measurement error (defined above). First, differentiation (also understood as variance attributable to "subjects") is achieved when the assessment design and implementation allow for a range of scores when they truly exist within and between candidates. Cases that result in ceiling or floor effects; raters who have difficulty differentiating between candidates, levels of performance, or dimensions of performance; and flawed rating tools can all reduce differentiation, and thus reliability. Second, measurement error can take many forms and is best-identified using generalizability theory.[55,64] Again, cases, raters, and items on a rating tool can all be sources of error. However, consistently the greatest source of error has been context specificity.[36,65-69] Context specificity refers to the finding that a candidate's performance on any one case is a poor predictor of performance on another. For example, context specificity would suggest that an individual's clinical performance when presented with a patient experiencing an asthmatic exacerbation would serve as a poor predictor of performance when presented with a patient suffering from sepsis. To improve predictions regarding future clinical performance, and perhaps to get closer to the candidates "true" level of performance, additional/multiple observations are required.[70]

Practice Implications—First, promote differentiation; the sample of cases used to assess candidates overall should challenge the candidate pool adequately, avoiding ceiling or floor effects while staying true to the construct of interest. Then, ensure raters have the ability to detect differences between levels and dimensions of performance and candidates, as poor rater performance may mask or mitigate differences. Similarly, ensure rating tools (described in more detail below) allow for differentiation. Again, poorly designed tools may mask differences that may otherwise be present.

Second, minimize error; sample performance broadly across a number of contexts/cases. Williams et al suggest seven to 11 observations are required to achieve a reasonable level of reliability;[47] however, this is only a useful starting point. Investigations of the effects of context specificity and other sources of error (eg, raters

and items) need to be conducted initially and on an ongoing basis since the degree of relative contribution by each source can vary by context/assessment features and offering, and different mitigation strategies may apply (eg, rater training).

Validity

Even the most reliable assessment processes would be limited without evidence of validity. Validity refers to the degree of confidence one has in the inferences made based on scores generated by an assessment process.[71-76] Because competence is an abstract construct, it can only ever be inferred (not directly measured) based on observations of behaviors in response to various clinical challenges/stimuli. In a practical sense, validity then refers to the accuracy or appropriateness of those inferences, including proposed interpretations (eg, degree of competence) or proposed use of assessment scores (eg, certification, remediation, and advancement). Importantly, validity is never actually achieved, rather different levels or degrees of evidence to support claims of validity are achieved; some are stronger than others. For example, demonstrating that a PBA is perceived by experts, candidates, and other stakeholders to be a suitable measure of clinical performance (ie, face validity) is a much weaker argument, or source of evidence, than having evidence that the same assessment actually predicts performance in future clinical settings (ie, predictive validity).[54] The stronger the collection of evidence in support of the inferences, the more one can have confidence in and defend decisions or interpretations made based on scores generated from an assessment process.

There are many threats to validity.[62] These can include poor reliability (discussed above), lack of authenticity, under- or over-construct representation, and/or construct irrelevant variance.[77] These threats may be present in the stimuli (eg, the clinical cases the candidates are presented with) or the response format (eg, the rating tool used and/or raters themselves).[77,78] First, authenticity refers to the degree to which the assessment context closely matches or aligns with future clinical contexts or performance expectations. Asking candidates to intubate a task trainer without any of the contextual forces that may be present in real clinical contexts (eg, prioritization, data gathering, and clinical reasoning) is a threat to validity because of the significant differences that exist between settings. In this example, poor authenticity results in larger extrapolations when making inferences regarding performance in real clinical contexts based on the assessment context.

A second and third threat to validity involves construct under- and over-representation.[72] When considering the stimulus and assuming a performance exam at the "shows how" or "does" level, construct under-representation refers to under sampling of the domain of possible encounters/construct of interest. To use a hypothetical example, if there are 100 different skills or patient types in the domain of possible encounters, and the sampling strategy only includes 10 of each, the risk of construct under-representation is higher than if 20 of each were included. Of course, not all elements of the construct can be included in a given assessment process. Therefore, one must apply a sampling strategy using a structured blue print and appropriate framework (one or more that define the construct of paramedic practice) to demonstrate and ensure adequate/appropriate representativeness. Construct over-representation (a form of construct irrelevant variance – described below) would be essentially the opposite, albeit less common, problem. That is, including content or behavior expectations in a PBA that is not representative of paramedic practice. Importantly, these concepts can apply to the cases designed/selected for an assessment process, but the rating tools as well. For instance, when developing or using a rating tool, if the measurement tool does not represent adequately the construct, by either missing important items or dimensions, or including items or dimensions that should not be, then both construct under- and over-representation are again possible, just in a different way.

Finally, another threat to validity is referred to as construct irrelevant variance and involves any systematic influences not directly related to the construct of interest.[79] This may include flawed cases, poor rating scales, inappropriate rating items, various forms of rater biases (eg, leniency and stringency or restriction of range), inadequate sampling, poor case difficulty (either too easy or too difficult), and unfamiliar equipment or simulators.[77] Any of these may artificially lower or elevate scores, causing scores to deviate inappropriately from the "true" score, thereby contributing error and threatening the confidence one would have that the scores generated lead to appropriate inferences/decisions (ie, validity claims).

Practice Implications—When assessment at the "does" level (ie, at the workplace, obviously the most authentic stimulus) is not feasible, simulations designed to replicate physical, conceptual, and emotional realism should be employed to support eventual extrapolations.[80] When considering the adaptive integration of multiple competencies, this extends to ensuring cases involve full clinical encounters consistent with the field of paramedicine, as opposed to fragmented or decontextualized skills which are arguably less authentic.

First, ensure the construct of paramedic practice is defined adequately to then inform sampling strategies (and case development) and mitigate the risk of under- and over- construct representation in PBA. Then, use a blueprint to make informed decisions regarding representativeness (ie, how well the domain profession-specific knowledge, skills, judgments, and possible encounters are included/excluded). The concept of construct representation applies similarly for rating tool selection and/or development.

Second, identify and eliminate potential or real sources of construct irrelevant variance (ie, noise) – features of the assessment that might influence scores other than the candidate's performance.

Multiple Observations Promote Both Reliability and Validity

The objective structured clinical evaluation (OSCE) aligns well with strategies to optimize reliability and validity. Originally developed in the 1970's in response to the low reliability associated with oral-based exams,[81,82] the OSCE has since taken on many variations across a variety of health professions and specialties[83-89] and has been adopted widely for high stakes licensing exams.[90-93] Objective structured clinical evaluations involve having candidates rotate through a series of standardized stations that collectively represent a larger specified construct of clinical competence (see Hodges for a detailed outline of how to design and build an OSCE).[80] The strength associated with OSCEs are in their ability to effectively address context specificity by including multiple stations, while simultaneously promoting construct representation by expanding the sample of content included in the assessment process. Despite the variations in the number of stations needed, a recent review found that after more than 30 years of use, the

OSCE produces reliable results,[94] has been reported as a gold standard for the assessment of competence in some settings,[95] and was found to be predictive of future clinical performance,[96] including paramedic contexts.[33]

Ruessler et al, in emergency medicine and Tavares et al, in paramedicine, provide two examples of PBA that are variations on the original OSCE.[33,97] In both instances, the main structure of timed, sequential rotations, standardization, ecological validity, construct representation, and multiple stations leading to multiple observations by multiple observers remain. However, rather than measure isolated skills or separated components of a complete workflow, both Ruessler and Tavares opted for more authentic or complete clinical cases within stations requiring the adaptive application of multiple competencies. While Ruessler included a combination of full clinical cases and more traditional OSCE stations (eg, ECG acquisition and interpretation or isolated intraosseous insertion), Tavares opted for full clinical cases in all stations. One other difference is that Ruesseler used station-specific checklists with weighted items, while Tavares used a generic un-weighted 7-dimension global rating scale (GRS)[34] across all stations (discussed in more detail below). Applying a generic rating tool allows researchers to assess each of the seven dimensions across a different context and rater, aggregating across stations rather than within stations. When considering only the full clinical cases in both the studies, reliabilities ranged from .55 (with five stations) to .79 (with six stations). Differences in reliability (even when using the same number of stations) may be associated with the rating tools, raters, the homogeneity of the group, and the cases used. For five of the seven dimensions, Tavares was able to demonstrate that performance in the simulation-based setting was associated with performance in real clinical contexts with real patients.[33]

Practice Implications—The OSCE and its variations serve as an effective SBA strategy based mainly on its inclusion of multiple observations by multiple raters across multiple contexts. This multiple sampling associated with OSCEs results in improved reliability by addressing context specificity and validity by promoting construct representation.

In paramedic contexts, as in emergency medicine, where interactions may be relatively brief and where the adaptive integration of multiple competencies is desired, validity may be optimized by using full clinical cases in each station (as opposed to isolated decontextualized skills).

Rating Tools

In making decisions regarding measurement tools, there are a number of factors to consider. For instance, if the goal is to focus on isolated procedural skills, and optimal performance of these skills is relatively linear, it may be more appropriate to employ some form of checklist where each step can be itemized. However, checklists may be limited when considering the integration of multiple competencies and may provide a false sense of objectivity.[98] However, selecting checklist purely for the pursuit of objectivity (defined as "a goal of measurement marked by freedom from subjective influences") or objectification, (defined as a "set of strategies designed to reduce measurement error"), researchers have identified no significant gains in reliability over GRSs.[98-101] Further, deconstructing clinical competence into its components parts, or assuming the linear accumulation of competencies results in accurate conclusions regarding competence, may be flawed. The challenges associated with checklists have led some to explore and adopt the use of GRSs. Global rating scales may have higher inter-station and inter-rater reliability, can be used across stations, and may better suited to capture "nuanced elements of expertise."[101-103]

Practice Implications—Both checklists and GRSs have their place. Global rating scales have emerged based on research that suggests: (a) not all that can be measured should be; (b) that competence may not be deconstructed effectively into component parts; (c) that the linear accumulation of competencies may not necessarily equal competence (as might be suggested by some checklists); (d) that GRSs may be more appropriate when considering complex practice; and that (e) "subjectivity" associated with GRS is no longer considered a bad word (described in more detail below).

Summary

As the paramedic profession expands in scope and contributions to the health care system, safety will be in large part dependent on the clinical competence of those granted access to the profession. The profession thus has a responsibility to ensure only those that are truly ready for independent practice are offered the public's trust. Doing this requires assessment strategies that optimize accuracy, trustworthiness, and defensibility when inferring competence based on behaviors exhibited in one context to future novel clinical contexts. In this narrative review, a number of foundational conceptual frameworks are highlighted, leading to practice implications aimed at meeting these assessment goals. In summary, an OSCE-like process (ie, using multiple stations/multiple views), including between seven and 11 simulation-based stations (as a starting point), involving full clinical cases and assessed using a GRS may promote defensibility when assessing paramedic clinical competence. The final number of stations can only be determined following a comprehensive blueprinting process using appropriate profession-specific frameworks, considerations of feasibility and logistical constraints, and a thorough statistical review using generalizability theory, or other methods, to identify sources of error, the data from which can be used to develop strategies aimed at improving psychometric and feasibility issues. The content for each station should include and require the adaptive integration of multiple competencies and maximum authenticity rather than isolated decontextualized skills. The performances can be evaluated using a GRS designed to represent the construct of paramedic clinical competence, and the scoring strategy can involve scoring across stations rather than within stations such that each relevant domain (eg, decision making) is evaluated across a variety of different contexts and raters, as opposed to passing or failing a station. Finally, any model must be followed with rigorous quality assurance/psychometric evaluation to determine and demonstrate evidence of both reliability and validity. A brief overview of these frameworks, along with the following emerging areas of research in assessment, which the paramedic community may consider monitoring, exploring further, and/or contributing to, is presented in Table 1.

Emerging Areas of Assessment Research

The assessment community continues to explore strategies aimed at optimizing the accuracy and utility of PBA. At a minimum, four areas of research have emerged as rich sources of study. These include: (a) a greater emphasis on programmatic

https://doi.org/10.1017/S1049023X15005166 Published online by Cambridge University Press

Guiding Conceptual Framework	Practice Implications for the Assessment of Paramedic Clinical Competence
Miller's Pyramid	No one-assessment strategy can capture all levels of Miller's Pyramid. This suggests the need for programmatic assessment with targeted strategies depending on areas of focus.[36] Assess at higher levels of Miller's pyramid what cannot be assessed more efficiently at lower levels. An underlying principle is that while performance at one level (eg, Level 2 "knows how") may be dependent on the level below it (ie, Level 1 "knows"), and that same level does not necessarily predict performance at the level above (ie, at Level 3 "shows how").[6,57] Competence involves a progression of knowledge, skills, and abilities, best measured in settings where the assessment context closely parallels the environment in which future clinical performance is expected to occur, mainly to minimize leaps in extrapolation in an inference based model.[40] Ideally, all clinical competence decisions would ultimately include assessment of behaviors exhibited in real clinical contexts with real patients (ie, at the "does" level).[57,58] However, in paramedic contexts in particular, WBAs are often associated with a number of challenges, such as lack of control over content (potentially leading to inappropriate case variability and/or complexity or a situation in which the sample of cases used for summative decisions may inadequately represent the clinical domain), interruptions due to patient safety concerns, and undue influence by many uncontrollable contextual factors.[59] As result of challenges associated with WBA, many have advocated for and adopted SBA;[59] ie, the "shows how" level) where ecological validity (ie, similarity between assessment and work-based settings), standardization, elimination of patient safety barriers, and complete control over content can be exercised. However, SBA only ever being a surrogate of reality requires evidence of both reliability and validity.
Reliability	Promote Differentiation: The sample of cases used to assess candidates overall should adequately challenge the candidate pool, avoiding ceiling or floor effects while staying true to the construct of interest. Ensure raters have the ability to detect differences between levels and dimensions of performance, and candidates as poor rater performance may mask or mitigate differences. Similarly, ensure rating tools allow for differentiation. Again, poorly designed tools may mask differences that may otherwise be present. Minimize Error: Sample performance broadly across a number of contexts/cases. Williams et al suggest seven to 11 observations are required to achieve a reasonable level of reliability;[47] however, this is only a useful starting point. Investigations of the effects of context specificity and other sources of error (eg, raters and items) need to be conducted initially and on an ongoing basis since the degree of relative contribution by each source can vary by context/assessment features and offering and different mitigation strategies may apply (eg, rater training).
Validity	When assessment at the "does" level (ie, at the workplace, obviously the most authentic stimulus) is not feasible, simulations designed to replicate physical, conceptual, and emotional realism should be employed to support eventual extrapolations.[80] When considering the adaptive integration of multiple competencies, this extends to ensuring cases involve full clinical encounters consistent with the field of paramedicine, as opposed to fragmented or decontextualized skills which are arguably less authentic. Ensure the construct of paramedic practice is adequately defined to then inform sampling strategies (and case development) and mitigate the risk of under- and over- construct representation in PBA. Use a blueprint to make informed decisions regarding representativeness (ie, how well the domain profession-specific knowledge, skills, judgments, and possible encounters are included/excluded). The concept of construct representation applies similarly for rating tool selection and/or development. Identify and eliminate potential or real sources of construct irrelevant variance (ie, noise) – features of the assessment that might influence scores other than the candidate's performance.
Multiple Sampling	The OSCE and its variations serve as an effective SBA strategy based mainly on its inclusion of multiple observations by multiple raters across multiple contexts. This multiple sampling associated with OSCEs result in improved reliability by addressing context specificity and validity by promoting construct representation. In paramedic contexts, as in emergency medicine, where interactions may be relatively brief and where the adaptive integration of multiple competencies is desired, validity may be optimized by using full clinical cases in each station (as opposed to isolated decontextualized skills).
Rating Tools	Both checklists and GRSs have their place. Global rating scales have emerged based on research that suggests (a) not all that can be measured should be; (b) that competence may not be effectively deconstructed into component parts; (c) that the linear accumulation of competencies may not necessarily equal competence (as might be suggested by some checklists); (d) that GRSs may be more appropriate when considering complex practice; and that (e) "subjectivity" associated with GRS is no longer considered a bad word.
Programmatic Assessments	Programmatic assessments should be considered in formative and summative assessment strategies. However, this in effect, moves the decision point from the point-in-time assessments to either individuals or committees who must somehow formulate decisions regarding competence, often with conflicting information. This is not inherently flawed; however it does require additional study to understand ways of addressing different defensibility challenges while remaining feasible.
Subjectivity in Assessments	Subjective assessments of paramedic clinical competence have value assuming that multiple views are obtained and processes for assembling and synthesizing the information have rigor.
Use of Narratives	Give that raters are viewed as naturally part of the assessment process, and the limitations associated with converting observations to rating tools, narratives in addition to or in place of numbers, may provide an important way of assessing competence.
Rater Cognition	While researchers continue to explore rater cognition, the emerging message seems to be to accept that objectivity in the assessment of clinical competence context may not exist and that rater judgment should be further understood, valued, and harnessed. There is growing discussion in the assessment literature and support for the value of rater judgment.[104] Rather than making attempts to objectify assessment processes, efforts should be made to align raters' natural thinking with assessment processes.[57,115,116] Assessment designers should embrace rater judgment and be less concerned about rater "subjectivity" when appropriate processes are place.

Table 1. Summary of Recommendations with Guiding Theoretical or Conceptual Framework
Abbreviations: GRS, global rating scale; OSCE, objective structured clinical examination; PBA, performance-based assessment; SBA, simulation-based assessment; WBA, work-based assessment.

https://doi.org/10.1017/S1049023X15005166 Published online by Cambridge University Press

assessment; (b) a movement away from a psychometric era to one that embraces the "subjective and collective;"[104,105] (c) the use of narratives in addition to, or in place of, rating scales;[106] and finally, (d) studies exploring rater cognition.[107,108] These relatively new areas of research are shaping the way the assessment community is viewing, implementing, and studying assessment practices. Reviewing each of these expanding bodies of literature in detail is beyond the scope of this report; however, brief summaries as they relate specifically to the growing field of PBA are provided.

Programmatic Assessments

Van der Vleuten and others suggest that assessments of individuals must encompass a variety of measures, carefully assembled such that a comprehensive understanding of the individuals degree of competence can be understood fully.[36] This includes different stimuli (eg, multiple-choice questions, simulation, and WBAs) and response formats throughout Miller's pyramid, but also within each level. Simulation and WBAs, for example, could be designed such that voids in WBA settings, based on the nature of the paramedic environment, could be complemented by SBA to contribute further to construct representation. Multiple sampling remains as an important feature, and strengths include having multiple sources of information usually from multiple perspectives and contexts collected longitudinally, each contributing a piece of the competence picture (when used for summative purposes). One of the methodological challenges that remain, however, with programmatic assessments of this kind includes how best to meaningfully assemble the data to form defensible decisions.

Practice Implications—Programmatic assessments should be considered in formative and summative assessment strategies. However, this, in effect, moves the decision point from the point-in-time assessments to either individuals or committees who must somehow formulate decisions regarding competence, often with conflicting information. This is not inherently flawed; however, it does require additional study to understand ways of addressing different defensibility challenges while remaining feasible.

Embracing the Subjective and Collective

Assessment of clinical competence has largely been dominated by a psychometric discourse, which is now being challenged.[36,105] The traditional psychometric view has been characterized by an unwavering pursuit of reliability and countless efforts to optimize "objectivity" while avoiding anything that might be classified as biased or subjective.[105] This review of the literature and recommendations above are reflective, in part, of these principles mainly because of what has dominated in the literature. However, researchers recently have begun to better understand some of the limitations of this psychometric discourse and have argued that much of what was once considered noise (ie, deviations from a "true" score), particularly where rater views are concerned, may actually be viewed as "signal" (ie, meaningful rich variation). The main philosophical shift being that each sample of information is meaningful, important, informed, and/or influenced by context and the rater, and when taken together, collectively contributes to the understanding of the candidate's readiness for independent practice. That is, there is no "true" score to capture, per se, rather an assembled co-constructed representation of the candidate's abilities. In other words, subjectivity may no longer be equated with unfairness, and rather than make attempts to optimize standardization, supposed objectivity, and control of biases, the assessment community is being encouraged to embrace subjectivity and the idiosyncrasies of raters. However, some caution a complete abandonment of psychometric principles in place of more constructivist methodologies.[109]

Practice Implications—Subjective assessments of paramedic clinical competence have value, assuming that multiple views are obtained and processes for assembling and synthesizing the information have rigor.

The Use of Narratives

This growing emphasis on the value of subjectivity, in addition to the complexity with which rater judgment occurs, has led the assessment community to consider the role of narratives. When raters observe, and then translate their observations to numbers on rating tools, a significant amount of information may be lost; the use of narratives may be a solution.[106] When used, the collection of narratives are then intended to be assembled and analyzed using qualitative strategies (eg, the identification of emerging themes in the data) to inform decisions or inferences regarding competence. The same concepts of validity apply, however, re-conceptualized to align with validity evidence that might be more consistent with constructivist perspectives or qualitative methodologies. Whereas in psychometric or positivist views, one of the questions researchers might ask may be "are the rater views the same enough" (promoting inter-rater reliability), in this emerging view, researchers may instead ask "are they different enough" (different enough to capture the construct in its entirety, as an example). Each RBA contributes only a piece of the larger construct, and as such, it is important that, collectively, the construct as defined by the profession is captured in some way, especially when narratives are unstructured. It is not entirely clear yet what the best method (taking into consideration feasibility) of assembling the narratives would be and/or how to work with the variable and interpretations that may arise.[110]

Practice Implications—Given that raters are viewed as naturally part of the assessment process and the limitations associated with converting observations to rating tools, narratives, in addition to or in place of numbers, may provide an important way of assessing competence.

Rater Cognition

Finally, researchers are placing the rater, rather than the process (eg, OSCE) or rating tools as the object of study.[107] This is a complicated area of research with many theoretical frameworks driving research questions, including the role of first impressions,[111] social contexts,[112] performance theories,[113] and cognitive load.[108] Briefly, researchers are finding that subjectivity in RBAs cannot be avoided, and that raters are anything but objective measurement tools reliably collecting and transferring information onto rating tools. For instance, Yeates et al find that, even when observing the same performance, raters attend to and give greater importance to different aspects of performance.[114] The assessment community now views raters as cognitive filters actively detecting and selecting information, processing the information while often influenced by a host of factors (eg, fatigue, previous performances, and inherent human capacity) and engaging in a translation process when forming categorical decisions or formulating

feedback. Further, raters may be engaging in internal strategies that satisfy, rather than optimize, their performance.

Practice Implication—While researchers continue to explore rater cognition, the emerging message seems to be to accept that objectivity in the assessment of clinical competence context may not exist and that rater judgment should be further understood, valued, and harnessed.

There is growing discussion in the assessment literature and support for the value of rater judgment.[104] Rather than making attempts to objectify assessment processes, efforts should be made to align raters' natural thinking with assessment processes.[57,115,116] Assessment designers should embrace rater judgment and be less concerned about rater "subjectivity" when appropriate processes are in place.

Conclusions

Assessment of clinical competence is a complex process requiring an appreciation of a broad and ever-expanding literature. A number of practice implications have been identified for the community to explore in more detail, and perhaps to integrate. Inherent in these practice implications is a (not so hidden) research agenda for the paramedic community to consider as the profession evolves. Those responsible for decisions regarding competence will need to engage in the scientific discussion to support safety in the professions through best practices in assessment. The conceptual frameworks discussed (eg, Kane's conceptualization of competence, Brennan's concept of a universe score, Miller's pyramid, reliability, validity, and multiple sampling) and the emerging programs of research (eg, programmatic assessment, harnessing inevitable "subjectivity," the use of narratives, and understanding how raters think) provide a foundation by which to structure, and perhaps to expand, assessments of paramedic competence. While this review has focused mainly on summative assessments at the entry to practice level, similar challenges exists in formative models, or even in the on-going maintenance of competence issue.

Acknowledgments

The authors would like to thank ORNGE Transport Medicine (Mississauga, Ontario, Canada), McMaster University (Hamilton, Ontario, Canada), Centennial College Paramedic Program (Toronto, Ontario, Canada), the Paramedic Association of Canada (Ottawa, Ontario, Canada), York Region Paramedic and Seniors Services (Sharon, Ontario, Canada), and the University of Ottawa (Ottawa, Ontario, Canada) for supporting this work.

References

1. Atack L, Maher J. Emergency medical and health providers' perceptions of key issues in prehospital patient safety. *Prehosp Emerg Care*. 2009;14(1):95-102.
2. Bigham BL, Buick JE, Brooks SC, Morrison M, Shojania KG, Morrison LJ. Patient safety in Emergency Medical Services: a systematic review of the literature. *Prehosp Emerg Care*. 2012;16(1):20-35.
3. Fairbanks RJ, Crittenden CN, O'Gara KG, et al. Emergency Medical Services provider perceptions of the nature of adverse events and near-misses in out-of-hospital care: an ethnographic view. *Acad Med*. 2008;15(7):633-640.
4. Jensen J, Croskerry P, Travers A. Consensus on paramedic clinical decisions during high-acuity emergency calls: results of a Canadian Delphi study. *Can J Emerg Med*. 2011;13(5):310-318.
5. Van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programs. *Med Educ*. 2005;39(3):309-317.
6. Miller G. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(9):S63-S67.
7. Holmboe E, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676-682.
8. Regener H. A proposal for student assessment in paramedic education. *Med Teach*. 2005;27(3):234-241.
9. Su E, Schmidt TA, Mann C, Zechnich AD. A randomized controlled trial to assess decay in acquired knowledge among paramedics completing a pediatric resuscitation course. *Acad Emerg Med*. 2000;7(7):779-786.
10. Lammers RL, Byrwa MJ, Fales WD, Hale RA. Simulation-based assessment of paramedic pediatric resuscitation skills. *Prehosp Emerg Care*. 2009;13(3):345-356.
11. Studnek JR, Fernandez AR, Shimberg B, Garifo M, Correll M. The association between Emergency Medical Services field performance assessed by high-fidelity simulation and the cognitive knowledge of practicing paramedics. *Acad Emerg Med*. 2011;18(11):1177-1185.
12. Unlüer EE, Yavasi O, Kara PH, et al. Paramedic-performed Focused Assessment with Sonography in Trauma (FAST) in the emergency department. *Ulus Travma Acil Cerrahi Derg*. [Turkish Journal of Trauma & Emergency Surgery: TJTES]. 2011;17(2):113-116.
13. Sahni R, Menegazzi JJ, Mosesso VN. Paramedic evaluation of clinical indicators of cervical spinal injury. *Prehosp Emerg Care*. 1997;1(1):16-18.
14. Zautcke JL, Lee RW, Ethington NA. Paramedic skill decay. *J Emerg Med*. 1987; 5(6):505-512.
15. LeBlanc V, MacDonald RD, McArthur B, King K, Lepine T. Paramedic performance in calculating drug dosages following stressful scenarios in a human patient simulator. *Prehosp Emerg Care*. 2005;9(4):439-444.
16. Hubble MW, Paschal KR, Sanders TA. Medication calculation skills of practicing paramedics. *Prehosp Emerg Care*. 2000;4(3):253-260.
17. Eastwood KJ, Boyle MJ, Williams B. Paramedics' ability to perform drug calculations. *West J Emerg Med*. 2009;10(4):240-243.
18. Rocca B, Crosby E, Maloney J, Bryson G. An assessment of paramedic performance during invasive airway management. *Prehosp Emerg Care*. 2000;4(2):164-167.
19. Ruetzler K, Roessler B, Potura L, et al. Performance and skill retention of intubation by paramedics using seven different airway devices—a manikin study. *Resuscitation*. 2011;82(5):593-597.
20. Manser T, Foster S, Gisin S, Jaeckel D, Ummenhofer W. Assessing the quality of patient handoffs at care transitions. *Qual Saf Health Care*. 2010;19(6):e44-e44.
21. LeBlanc VR, Regehr C, Tavares W, Scott AK, MacDonald R, King K. The impact of stress on paramedic performance during simulated critical events. *Prehosp Disaster Med*. 2012;27(4):369-374.
22. Gordon DL, Issenberg SB, Gordon MS, LaCombe D, McGaghi WC, Petrusa ER. Stroke training of prehospital providers: an example of simulation-enhanced blended learning and evaluation. *Med Teach*. 2005;27(2):114-121.
23. Losek JD, Szewczuga D, Glaeser PW. Improved prehospital pediatric ALS care after an EMT-paramedic clinical training course. *Am J Emerg Med*. 1994;12(4):429-432.
24. Pointer JE. Clinical characteristics of paramedics' performance of pediatric endotracheal intubation. *Am J Emerg Med*. 1989;7(4):364-366.
25. Lee KHK, Grantham H, Boyd R. Comparison of high-and low-fidelity mannequins for clinical performance assessment. *Emerg Med (Fermantle)*. 2008;20(6):508-514.
26. Wyatt A, Fallows B, Archer F. Do clinical simulations using a human patient simulator in the education of paramedics in trauma care reduce error rates in preclinical performance? *Prehosp Emerg Care*. 2004;8(4):435-436.
27. Castle N, Owen R, Hann M, Naidoo R, Reeves D. Assessment of the speed and ease of insertion of three supraglottic airway devices by paramedics: a manikin study. *Emerg Med J*. 2010;27(11):860-863.
28. Hein C, Owen H, Plummer J. A training program for novice paramedics provides initial laryngeal mask airway insertion skill and improves skill retention at 6 months. *Simul Healthc*. 2010;5(1):33-39.
29. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education. *JAMA*. 2011;306(9):978-988.
30. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009;302(12):1316-1326.
31. Martin M, Hubble MW, Hollis M, Richards ME. Inter-evaluator reliability of a mock paramedic practical examination. *Prehosp Emerg Care*. 2012;16(2):277-283.
32. Janing J, Sime WR. Inter-rater reliability of paramedic student field performance evaluations. *Prehosp Emerg Care*. 1999;3(3):265-266.
33. Tavares W, LeBlanc VR, Mausz J, Sun V, Eva KW. Simulation-based assessment of paramedics and performance in real clinical contexts. *Prehosp Emerg Care*. 2014; 18(1):116-122.
34. Tavares W, Boet S, Theriault R, Mallette T, Eva KW. Global rating scale for the assessment of paramedic clinical competence. *Prehosp Emerg Care*. 2012;17(1):57-67.
35. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Edu Theory Pract*. 1996;1(1):41-67.
36. Van der Vleuten C, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol*. 2010;24(6):703-719.

37. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology*. 2010;112(4):1041-1052.
38. Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Can J Anaesth*. 2012;59(2):182-192.
39. Hodges B. OSCE! Variations on a theme by Harden. *Med Educ*. 2003;37(12): 1134-1140.
40. Hodges B. Validity and the OSCE. *Med Teach*. 2003;25(3):250-254.
41. Hodges BD. The objective structured clinical examination: three decades of development. *J Vet Med Educ*. 2006;33(4):571-577.
42. Norcini JJ, McKinley DW. Assessment methods in medical education. *Teach Teach Educ*. 2007;23(3):239-250.
43. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206-214.
44. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357(9260):945-949.
45. Epstein R, Hundert E. Defining and assessing professional competence. *JAMA*. 2002;287(2):226-235.
46. Epstein R. Assessment in medical education. *N Engl J Med*. 2007;356(4):387-396.
47. Williams R, Klamen D, McGaghie W. Special Article: cognitive, social, and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270-292.
48. Swanson D, Norman G, Linn R. Performance-based assessment: lessons from the health professions. *Educ Res*. 1995;24(5):5.
49. Regehr G, Eva K, Ginsburg S, Halwani Y, Sidhu R. Assessment in postgraduate medical education: trends and issues in assessment in the workplace. *The Future of Medical Education in Canada*. 2011.
50. Schuwirth L, Southgate L, Page GG, et al. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ*. 2002; 36(10):925-930.
51. Kane MT. The assessment of professional competence. *Eval Health Prof*. 1992; 15(2):163-182.
52. Frank JR, Mungroo R, Ahmad Y, Wang M, DeRossi S, Horsley T. Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Med Teach*. 2010;32(8):631-637.
53. Tavares W, Mausz J. Assessment of non-clinical attributes in paramedicine using multiple mini-interviews. *Emerg Med J*. 2013;32(1):70-75.
54. Tavares W, et al. Simulation-based assessment of paramedics and performance in real clinical contexts. *Prehosp Emerg Care*. 2014;18(1):116-122.
55. Brennan RL. *Generalizability Theory*. Heidelberg, Berlin: Springer Verlag; 2001.
56. Brennan RL. *Educational Measurement*. Westport, Connecticut USA: Praeger Pub Text; 2006.
57. Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ*. 2012;46(1): 28-37.
58. Norcini JJ. Current perspectives in assessment: the assessment of performance at work. *Md Educ*. 2005;39(9):880-889.
59. Boulet J, Swanson D. Psychometric challenges of using simulations for high-stakes assessment. *Simulations in Critical Care Education and Beyond*. Des Plains, Illinois USA: Society of Critical Care Medicine; 2004: 119-130.
60. Eva K. "Assessment Strategies in Medical Education." In: Salerno-Kennedy, (ed). *Medical Education: State of the Art*. Halifax, Nova Scotia, Canada: Nova Scotia Publishers, Inc.; 2010.
61. Downing S. Reliability: on the reproducibility of assessment data. *Med Educ*. 2004; 38(9):1006-1012.
62. Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*. 2004;23(1):17-27.
63. Streiner D, Norman G. *Health Measurement Scales: A Practical Guide to their Development and Use*, Fourth ed. New York, New York USA: Oxford University Press; 2008.
64. Brennan R. Performance assessments from the perspective of generalizability theory. *Appl Psychol Meas*. 2000;24(4):339-354.
65. Eva KW. On the generality of specificity. *Med Educ*. 2003;37(7):587-588.
66. Eva KW, Neville AJ, Norman GR. Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Acad Med*. 1998; 73(10):S1-S5.
67. Van der Vleuten C, Swanson D. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med*. 1990;2(2):58-76.
68. Newble D, Swanson D. Psychometric characteristics of the objective structured clinical examination. *Med Educ*. 1988;22(4):325-334.
69. Van der Vleuten CP. When I say… context specificity. *Med Educ*. 2014; 48(3):234-235.
70. Downing S, Yudkowsky R. *Assessment in Health Professions Education*. New York, New York USA: Taylor & Francis; 2009.
71. Kane MT. "Validity." In: *Educational Measurement*. Brennan BL, (ed). Westport, Connecticut USA: Praeger Pub. Text; 2006.
72. Downing S. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837.
73. Schuwirth LW, Van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48.
74. Kane M. Current concerns in validity theory. *Journal of Educational Measurement*. 2001;38(4):319-342.
75. Kane M. Validating score interpretations and uses. *Language Testing*. 2012; 29(1):3-17.
76. Kane MT. An argument-based approach to validity. *Psychol Bull*. 1992; 112(3):527-535.
77. Downing S, Haladyna T. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327-333.
78. Downing S. Threats to the validity of clinical teaching assessments: what about rater error? *Med Educ*. 2005;39(4):353-355.
79. Haladyna T, Downing S. Construct irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practice*. 2004;23(1):17-27.
80. Hodges B, Hanson M, McNaughton N, Reghr G. Creating, monitoring, and improving a psychiatry OSCE: a guide for faculty. *Acad Psychiatry*. 2002; 26(3):134-161.
81. Harden R, et al. Assessment of clinical competence using objective structured examination. *Br Med J*. 1975;1(5955):447-451.
82. Harden R. What is an OSCE? *Med Teach*. 1988;10(1):19-22.
83. Turner J, Dankoski M. Objective structured clinical exams: a critical review. *Fam Med*. 2008;40(8):574-578.
84. Hatala RM, Stevenson M, Downie WW, Wilson GM. Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. *BMC Med Educ*. 2011;11(1):23.
85. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini interview. *Med Educ*. 2004;38(3):314-326.
86. Ryan S, Stevenson K, Hassell AB. Assessment of clinical nurse specialists in rheumatology using an OSCE. *Musculoskeletal Care*. 2007;5(3):119-129.
87. Rushforth HE. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. *Nurse Educ Today*. 2007; 27(5):481-490.
88. Park RS, Chibnall JT, Blaskiewicz RJ, Furman GE, Powell JK, Mohr CJ. Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Acad Psychiatry*. 2004;28(2):122-128.
89. Manogue M, Brown G. Developing and implementing an OSCE in dentistry. *Eur J Dent Educ*. 1998;2(2):51-57.
90. Reznick R, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Acad Med*. 1996;71(1):S19-S21.
91. Smee SM, Dauphinee WD, Blackmore DE, Rothman AI, Reznick RK, Des Marchais J. A sequenced OSCE for licensure: administrative issues, results and myths. *Adv Health Sci Edu Theory Pract*. 2003;8(3):223-236.
92. Lee YS. OSCE for the medical licensing examination in Korea. *The Kaohsiung J Med Sci*. 2008;24(12):646-650.
93. Boulet JR, Smee SM, Dillon GF, Gimpel JR. The use of standardized patient assessments for certification and licensure decisions. *Simul Healthc*. 2009;4(1):35-42.
94. Patrício MF, Juliao M, Fareleira F, Carniero AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Med Teach*. 2013; 35(6):503-514.
95. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination (OSCE). The new gold standard for evaluating postgraduate clinical performance. *Ann Surg*. 1995;222(6):735-742.
96. Tamblyn R, Abrahamowicz M, Brailovsky C, et al. Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA*. 1998;280(11):989-996.
97. Ruesseler M, Weinlich M, Byhahn C, et al. Increased authenticity in practical assessment using emergency case OSCE stations. *Adv Health Sci Edu Theory Pract*. 2010;15(1):81-95.
98. Vleuten C, Norman G, Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ*. 1991;25(2):110-118.
99. Norman G, Vleuten C, Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency, and acceptability. *Med Educ*. 1991;25(2):119-126.
100. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73(9):993-997.
101. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49(2):161-173.

102. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med*. 1999;74(10):1129-1134.
103. Malau-Aduli BS, Mulcahy S, Warnecke E, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. *Creat Educ*. 2012;3(6A):937-942.
104. Eva KW, Hodges BD. Scylla or Charybdis? Can we navigate between objectification and judgment in assessment? *Med Educ*. 2012;46(9):914-919.
105. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
106. Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol*. 2013;21(4):668.
107. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48 (11):1055-1068.
108. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Edu Theory Pract*. 2013;18(2):291-303.
109. Norman G. When I say...reliability. *Med Educ*. 2014;48(9):946-947.
110. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296-306.
111. Wood TJ. Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Edu Theory Pract*. 2014;19(3):409-427.
112. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(10):S1-S7.
113. Govaerts MJ, Van de Wiel MW, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Edu Theory Pract*. 2013;18(3):375-396.
114. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently. *Adv Health Sci Edu Theory Pract*. 2013;18(3):325-341.
115. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med*. 2010;85(5):780.
116. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace based assessment scales. *Med Educ*. 2011;45(6):560-569.