


Applying Kane’s validity framework to a simulation based assessment of clinical competence

Walter Tavares^{1,2,3}  · Ryan Brydges¹ · Paul Myre⁵ · Jason Prpic⁵ · Linda Turner⁶ · Richard Yelle⁴ · Maud Huiskamp⁶

Received: 22 February 2017 / Accepted: 22 October 2017 / Published online: 27 October 2017
© Springer Science+Business Media B.V. 2017

Abstract Assessment of clinical competence is complex and inference based. Trustworthy and defensible assessment processes must have favourable evidence of validity, particularly where decisions are considered high stakes. We aimed to organize, collect and interpret validity evidence for a high stakes simulation based assessment strategy for certifying paramedics, using Kane’s validity framework, which some report as challenging to implement. We describe our experience using the framework, identifying challenges, decisions points, interpretations and lessons learned. We considered data related to four inferences (scoring, generalization, extrapolation, implications) occurring during assessment and treated validity as a series of assumptions we must evaluate, resulting in several hypotheses and proposed analyses. We then interpreted our findings across the four inferences, judging if the evidence supported or refuted our proposed uses of the assessment data. Data evaluating “Scoring” included: (a) desirable tool characteristics, with acceptable inter-item correlations (b) strong item-total correlations (c) low error variance for items and raters, and (d) strong inter-rater reliability. Data evaluating “Generalizability” included: (a) a robust sampling strategy capturing the majority of relevant medical

Electronic supplementary material The online version of this article (<http://doi.org/10.1007/s10459-017-9800-3>) contains supplementary material, which is available to authorized users.

✉ Walter Tavares
walter.tavares@utoronto.ca

¹ The Wilson Centre, Department of Medicine, University of Toronto/University Health Network, 200 Elizabeth Street, 1ES-565, Toronto, ON M5G 2C4, Canada

² Post-MD Education (Post-Graduate Medical Education/Continued Professional Development), University of Toronto, Toronto, ON, Canada

³ Paramedic and Senior Services, Community and Health Services Department, Regional Municipality of York, Newmarket, ON, Canada

⁴ Ornge Transport Medicine, Base Hospital and Clinical Affairs, Mississauga, ON, Canada

⁵ Health Sciences North Base Hospital, Sudbury, ON, Canada

⁶ Sunnybrook Base Hospital, Toronto, ON, Canada

directives, skills and national competencies, and good overall and inter-station reliability. Data evaluating “Extrapolation” included: low correlations between assessment scores by dimension and clinical errors in practice. Data evaluating “Implications” included low error rates in practice. Interpreting our findings according to Kane’s framework, we suggest the evidence for scoring, generalization and implications supports use of our simulation-based paramedic assessment strategy as a certifying exam; however, the extrapolation evidence was weak, suggesting exam scores did not predict clinical error rates. Our analysis represents a worked example others can follow when using Kane’s validity framework to evaluate, and iteratively develop and refine assessment strategies.

Keywords Assessment · Competence · OSCE · Paramedic · Simulation · Validation · Validity

Introduction

Safe and effective healthcare requires rigorous assessment of clinicians’ ability to deliver care. Conducting accurate and trustworthy assessments can be challenging given the complexity associated with clinical competence in the health professions. For instance, domains of competence are ever growing (Frank et al. 2014; Tavares et al. 2016) and some suggest assessments must be sensitive to complex performance requirements, such as the ability to adaptively and flexibly integrate selective competencies. (Myolopolous and Regehr 2011) Further, competence is abstract and not directly measurable, requiring inferences to be made based on how candidates behave in response to clinical stimuli. (Kane 1999) Given the proliferation of assessment tools/processes in health professions education, there is a need to organize the evidence on how well these reflect and assess proposed competencies and support decision-making regarding clinician ability. Validity frameworks for evaluating assessment tools/processes provide conceptual and practical guidance for collecting, organizing, analyzing, and evaluating the resulting evidence. When applied appropriately these frameworks, ultimately, help educators use assessments to make appropriate and justified decisions. (Kane 2013a, b).

Many validity frameworks exist for evaluating assessment practices, each offering unique yet overlapping views and processes. (St-Onge et al. 2017) Common among them is the idea that validity pertains to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test score. (Messick 1989) Kane’s validity framework, (Kane 2013a, b) builds on the earlier work of Messick, proposing a network of inferences and assumptions inherent to any assessment process that may threaten decisions or conclusions. According to Kane, assessors must identify, state and test their inferences to determine if the evidence supports or refutes the decisions they make based on assessment scores or narratives. (Kane 2013a, b) Kane proposes that assessors collect evidence across four inferences (as appropriate): scoring, generalization, extrapolation, and implications. Specifically, assessment activities are evaluated on whether assessors generate appropriate and consistent scores after observing performance (i.e., scoring), on how well the observed sample represents the broader range of possible performances (i.e., generalization), on whether the observed performance relates to performance in the real world (i.e., extrapolation), and on how the assessors’ decision about performance, such as pass or fail, impacts the individual, the training program, the profession and/or society (i.e., implications).

Kane outlines numerous steps for validation of assessment tools/processes. Once educators define their construct of interest, such as 'communication' or 'specialist X', they must specify how they will interpret and use the assessment scores (e.g., to judge transition from junior to senior training levels). Next, as they develop their assessment plan, they must (a) identify and explicitly state their assumptions (e.g., that raters can consistently distinguish the communication skills of junior and senior trainees), (b) prioritize the weakest and most questionable assumptions as targets for analysis, and (c) develop a plan for collecting evidence to test each assumption. (Cook and Hatala 2016) Kane calls the resulting list of claims the "interpretation-use argument". After collecting the evidence, educators compare what they found to their original assumptions, and develop a final 'validity argument' describing whether their assumptions are supported or refuted, where important gaps remain (if any), and what corrections might be necessary.

While Kane's validity framework has translated into new ways of organizing validation efforts, a recent review highlights how validity is still defined broadly and inconsistently, and how a tension exists between conceptual recommendations and actual validation practices. (St-Onge et al. 2017) Despite guides to help educators apply Kane's validity framework, (Cook and Hatala 2016; Clauser et al. 2012) few have used it to evaluate assessment processes in health profession education. Application continues to be challenging, with those conducting validation studies reporting difficulty using such guides. (Hatala et al. 2015; Ponton-Carass et al. 2016; Cook et al. 2014) Others have noted more broadly that validation remains poor, that few give it the attention it deserves, that few list their claims, and that fewer still evaluate those claims. (Brennan 2013) Yet the community continues to recognize the need to provide such evidence.

We followed Kane's framework to produce a worked example of how it can be used to conduct a validation study of a simulation-based assessment. However, we anticipated our efforts to translate this conceptual framework into pragmatic reality would be marked by challenges. In the sections below, we describe our setting and assessment strategy, define our interpretation-use argument, specify our assumptions and the evidence we collected to test them, and contrast the results of our assessment process with our original argument. We also reflect on and discuss how we applied the framework throughout the various stages and discuss any tensions we worked to resolve. Using a high stakes assessment process, we demonstrate how to collect empirical data and report our *judgment processes* which were essential to this validation effort. Our 'worked example' is designed to highlight how the framework can be used to direct subsequent improvements to an assessment process while also providing empirical data for the assessment process. Moreover, we present our example in a way that allows readers to find similarities or differences with their own assessment and validation frameworks and processes.

Methods

We applied Kane's validity framework when collecting and evaluating evidence for a performance based assessment in a paramedic context. In doing so we had to work through several uncertain points of 'knowledge translation'. For example, we had to decide which inferences to prioritize (i.e., implications of assessment) and had to judge the resulting evidence as favourable, or not, with no explicit guidelines. We monitored, documented and reflected on the challenges, struggles and ultimately the decisions we made in applying theory to practice. What follows is a description of our approach, including analysis, results

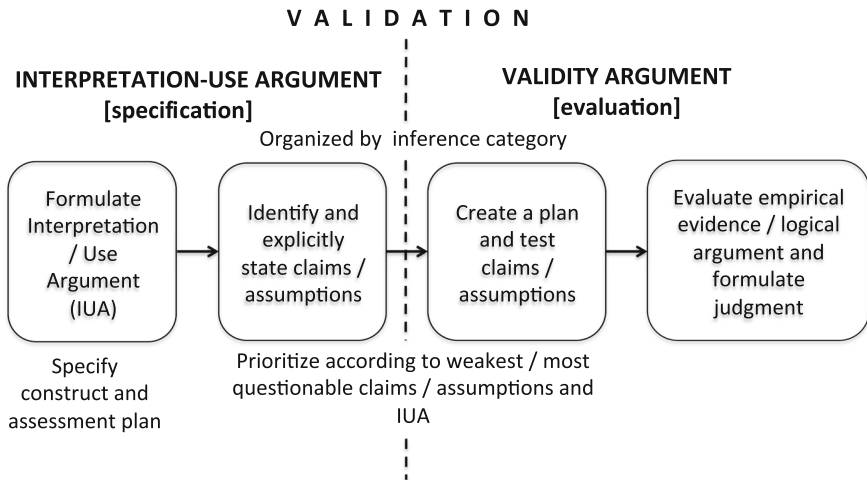
and interpretations of this particular performance based assessment, which we intend to serve as a type of worked example. We then use this example and our reflections in our discussion to highlight the challenges and lessons learned when applying Kane’s framework.

Study overview—applying Kane’s validity framework

We interpreted and operationalized Kane’s conceptual framework into the process illustrated in Fig. 1. Our construct of interest was primary care paramedic competence at the entry to practice level. We designed a seven station objective structured clinical examination (OSCE) involving full clinical cases relevant to paramedic practice. We planned that educators would use test scores to inform decisions regarding readiness for entry to paramedic practice at the primary care level and to predict candidates’ performance in subsequent clinical settings. Based on this interpretation and use, we articulated the many assumptions inherent in the assessment process. We organized our assumptions according to Kane’s four inference categories: scoring, generalization, extrapolation, and implications. Our list of assumptions became our interpretation-use argument (abbreviated below as IUA), which we used to construct our hypotheses and associated analyses, shown in Table 1. After we collected all the pre-defined validity evidence, we compared our findings to our original assumptions and developed a final ‘validity argument’ describing whether the assumptions were supported or not, where important gaps remain, and what corrections to the assessment process might be necessary.

Participants/candidates

In Ontario Canada paramedics are authorized to provide care under the delegating authority of a medical director and Base Hospital program. Once candidates complete their



Inference categories = scoring; generalization; extrapolation; implications

Fig. 1 Schematic illustrating the process of validation, including specification of the inherent claims/ assumption associated with the interpretation-use argument (from left to right, boxes 1 and 2) and evaluation of those claims (boxes 3 and 4)

Table 1 Summary of inference categories, definition for each, appraisal of existing evidence, rationale for prioritization, hypotheses generated to test our assumptions, and associated analyses. Our study design was informed by our prioritization of assumptions given our intended interpretation/use

	Definition	Assumptions/ hypotheses	Appraisal of existing evidence	Why prioritized in this sequence?	Analyses
Implications	Refers to the process of moving from scores to decisions about the individuals. (Kane 2013a, b)	<i>Hypothesis 1:</i> Applying the borderline group method will produce low failure rates given our homogenous and highly selected participants. <i>Hypothesis 2:</i> Error rates will be similar, if not lower, than previous assessment processes. That is, we expected the cohort that passed the assessments would not lead to more performance errors in practice than previously recorded.	Evidence suggests a borderline group method establishes pass-fail rates consistent with the level of learner ability. (Tavares et al., 2012)	Need robust implications evidence to support our "decision rules" for making inferences of competence or readiness for independent practice.	Using descriptive and correlational and ANOVA procedures we tested if (a) we could make predictions and (b) whether clinical error rates were considered high and better or worse than earlier cohorts who were assessed in a different way.
Extrapolation	Refers to evidence for how well candidates will perform in future and novel clinical contexts (what Kane refers to as the target domain). (Kane 2013a, b)	<i>Hypothesis 1:</i> Our assessment process will predict (or be associated with) clinical performance in future real world clinical contexts.	Evidence shows GRS scores discriminate levels of expertise and correlate well with workplace based assessments. (Tavares et al., 2012)	Claim in our IUA that exam scores can predict performance in the clinical setting, however, clinical errors have not been studied as a correlate previously.	We analyzed the correlation between assessment scores across seven dimensions and clinical errors in practice over a 6-month period immediately following the assessment process.

Table 1 continued

	Definition	Assumptions/ hypotheses	Appraisal of existing evidence	Why prioritized in this sequence?	Analyses
Generalization	Refers to the degree to which the assessment protocol (e.g., selected stations and items) represents all of the theoretically possible clinical events. (Kane 2013a, b)	<p><i>Hypothesis 1:</i> Our sampling strategy was sufficient to establish a reliable measure of participants' paramedicine competencies.</p> <p><i>Hypothesis 2:</i> Scores obtained in this setting would be similar if an entirely new set of cases were to be used as evidenced by our generalizability theory (GT) analyses. (Brennan 2001)</p>	Previous decision study suggested 10 stations required for simulation-based assessment. (Tavares et al., 2014)	Logistics required that we choose seven stations. Therefore we needed to evaluate this to ensure the number of stations and sampling strategy was appropriate.	We compared our assessment maps to national and provincial standards and guidelines. We conducted GT analyses which involves using ANOVA to partition variances to various facets (e.g., subject, stations, raters, items and their interactions) included in the assessment process. (Brennan 2001)
Scoring	Refers to the process of moving from an observed performance to an observed score; it includes the scoring rules, rubric and scoring procedures. (Kane 2013a, b)	<p><i>Hypothesis 1:</i> Dimensions on the GRS would demonstrate evidence of independence (i.e., low to moderate inter-item correlations).</p> <p><i>Hypothesis 2:</i> Reliability analyses would reveal low error variance for items and/or raters.</p> <p><i>Hypothesis 3:</i> Inter-rater reliability would be moderate to high.</p>	The scoring rubric has been established using rigorous methods, with supportive evidence for scoring items. Satisfactory inter-rater reliability but in a narrower context, (Tavares et al., 2012) suggesting a need for continued evaluation.	Given previous evidence, scoring assumed to be established, though studies can always continue to evaluate such evidence, especially where GT analyses can be conducted.	Using correlations and the GT analyses described above, we calculated data for item analyses including inter-item and item-total correlations as well as reliability analyses.

training, those selected for employment are provisional, pending Base Hospital clinical competence screening and certification. All candidates had completed paramedic training and were seeking Base Hospital certification to complete employment eligibility requirements.

Procedures

Study setting—simulation-based assessment of paramedicine competencies

The Ontario Base Hospital clinical performance assessment is a simulation based entry to practice (i.e., high stakes) exam including seven independent standardized stations in an OSCE format (i.e., candidates rotate from station to station). Each station involved a full clinical case consistent with paramedic practice, from initial patient contact in variable contexts to implementation of treatment plans, transport or transfer of care. Paramedic candidates moved sequentially through each station until all seven were completed. Each station included a standardized patient or mannequin, a standardized “partner” (an actor, certified and experienced as a PCP), various props and other contextual cues as the case required. Candidates were expected to lead the patient interaction, conduct assessments and interpret findings, formulate care plans, use the “partner” and other resources as needed until time expires. Standardized “partners” were programmed to function effectively and efficiently but not to contribute to care pathways without direction from the candidate. Each station was scheduled for 12 min regardless of performance or interventions and standardized so that candidates experienced the same case, bystander and partner performances. No feedback was provided immediately following the performance. Candidates were given 5 min to rest between stations.

Case/content development

A content committee made up of eight Base Hospital program coordinators (representing all of Ontario) was assembled to prepare the cases. Content was blueprinted using documents defining or informing paramedic practice in Ontario including: (a) Ontario Advanced Life Support patient care standards (ALS-PCS); (Ontario Ministry of Health Emergency Health Services Branch 2007) (b) a skills and prioritized disease classification list derived from these standards; (c) the National Occupational Competency Profile (NOCP); (Paramedic Association of Canada 2016) (d) a patient profile list (e.g., age groups and types). Basic Life Support (BLS) skills (e.g., fracture management) (Ontario Ministry of Health Emergency Health Services Branch 2007) (b) were not blueprinted given the Base Hospital's role of medical oversight of advanced life support, but were not excluded to promote authenticity. Clinical cases, derived from actual patient encounters, were created including contextual cues, patient characteristics and behaviors, medical history, case progression and performance expectations by dimension included on the rating tool (described below) and mapped to these documents. These were reviewed and revised until consensus was reached on all details. The final set of cases, performance expectations and blueprint were then reviewed and approved by all members of the assessment committee prior to implementation. The assessment spanned 6 days. The same cases were used with only changes to surface characteristics (e.g., the setting in which the patient was found; home, public space).

Raters and rater training

Each station was assigned two raters. Raters were selected from across Ontario and represented active paramedics who were also Base Hospital educators and Base Hospital staff with paramedic credentials. All raters were provided with a six-hour orientation session three weeks prior to the assessment date. The orientation included content describing the conceptual framework for the assessment design (e.g., rationale for broad sampling, importance of authenticity), the role of the rater (e.g., differentiate between candidates and dimensions, contribute data for the purposes of pass/fail decisions), a detailed description of the rating tool (e.g., scale development, characteristics) and how to apply it (e.g., adopting the definitions, isolating dimensions of performance). Modified forms of performance dimension and frame of reference training (Roch et al. 2011; Woehr and Huffcutt 1994) were also included as part of the orientation. However, there was no attempt to formally calibrate raters, rather, simply to understand and apply supportive rating behaviors/rules (e.g., an observed behavior may apply to more than one dimension). Adopting the performance expectations outlined by the case writers was also emphasized.

Scoring and standard setting procedures

Two raters per station scored candidates' performances using the same paper-based seven-dimension global rating scale (GRS) (described in more detail below). In this model, raters were nested in cases (i.e., any two paired raters remained in the same station for all candidates for a given assessment session). Raters were instructed to avoid sharing views on any performance(s) or sharing scores assigned, to avoid rater calibration over time, and to allow for calculations of inter-rater reliability relevant to future practice. Scores were recorded directly onto the paper forms and transferred to an electronic database for analysis.

Rather than calculating candidates' mean score by station, we calculated their mean score for each dimension across the seven stations. Using the same rating tool across all stations—generic dimensions applicable to all cases—allowed us to calculate scores in both directions (within and across stations). In choosing to explore the analysis across stations, the interpretation is that we now get to observe “Situation Awareness” for example, across seven contexts across and seven raters, rather than have it blended into a within-station score. Using a non-compensatory scoring model allowed us to use each of these dimensions (intended to represent paramedic practice) individually. Scores on a seven-point scale were converted to percentages and standard error of measurement (SEM) was calculated for scores on each dimension. We then calculated cut scores for each dimension using a borderline group method (Humphrey-Murto and MacFadyen 2002). A non-compensatory scoring model (i.e., candidates were required to achieve above the cut score on all dimensions to be deemed successful) was used as the decision rule regarding pass-fail decisions.

Data collection/measurement tools

To evaluate the candidate's clinical performance we used a previously developed seven-dimension (situation awareness, history gathering, patient assessment, decision-making, resource utilization, communication and procedural skill) global rating scale (GRS), designed specifically for the assessment of paramedics. (Tavares et al. 2012, 2014) Each

dimension includes a definition and is scored using a seven-point adjectival scale. The seven points are anchored with labels and definitions that make reference to safety, performance standards and level of supervision and/or suitability for progression. This GRS has been applied in similar ways previously with evidence suggesting a rigorous approach to establish the scoring rubric, satisfactory inter-rater reliability, that the tool can discriminate levels of expertise, that simulation-based assessment scores correlate well with scores on workplace based assessments, and that use of a cut-score failed more novice students than entry-to-practice and experienced paramedics. (Tavares et al. 2014).

As our IUA specifies informing decisions regarding readiness for entry to paramedic practice at the primary care level and predicting performance in real clinical settings we sought and were obligated to identify additional measurement outcomes. This meant gathering data on clinical errors once in practice. Clinical errors in this context were defined as instances in which paramedics deviate from established best practices or medical directives (e.g., drug administration when clinical parameters are not met, dosing errors, prolonged arrhythmia without intervention) as reflected on electronic patient care records or through self-reporting. (Sunnybrook Centre for Prehospital Medicine Regional Base Hospital 2016) Monitoring for these errors is part of ongoing quality assurance using a custom-built in-house software program that scans and filters electronic patient care records using predefined patient care algorithms. Quality assurance for the software itself is established by crosschecking algorithms and scripts in advance and on an ongoing basis. Those records identified by the software as errors are then reviewed by quality assurance personnel who apply standardized audit tools to make final determinations regarding actual presence of errors. Clinical errors are categorized as minor, moderate and critical referring to little, moderate or high potential for adversely affecting patient outcomes respectively. First, error rates were obtained from historical databases, which served as a baseline. We retrieved data from a random selection of individuals who were assessed using a different and earlier model. Second, error rates were tallied for each individual in this cohort of candidates regardless of degree for 210 days post assessment date overall and by taking into consideration variable call volumes/patient contacts for the same time period. Selecting 210 days was an arbitrary length of time, but sufficiently lengthy to establish a reasonable data set. Our intention was not to compare assessment processes directly (described below), but to have data by which to explore sudden changes in error rates as a result of this new assessment method.

Analysis plan

Our hypotheses and corresponding analysis plans are provided in Table 1. We organized our analyses by inference category in the following way: implications, extrapolation, generalization and scoring. We prioritized our analysis in this way based first on our IUA, and second, on what were perceived to be our most questionable/weakest assumptions (includes considering existing evidence) when trying to formulate a validity argument. As described above, this drove our data collection and analysis plan. In total we identified and chose to examine eight assumptions that could help us formulate an argument in support of the inferences and decisions we make based on the scores generated. For example, given that our IUA indicated intentions to use scores to inform decisions regarding readiness for entry to paramedic practice and to predict performance, it was important to test specifically assumptions informing implications (i.e., the process of moving from scores to decisions about the individuals). and extrapolation (i.e., evidence for how well candidates will perform in future and novel clinical contexts). We include scoring and generalization data

and analyses to further support and build the validity argument. Kane's validity framework emphasizes a chain of inferences from the generation of scores to decisions regarding test takers, a chain that can be conceptualized as a series of "bridges that must be crossed" before sufficient evidence (existing or new) is achieved. (Kane 2012) As such, we report our results guided by this step-by-step conceptualization.

Results

A total of 125 candidates participated in the exam, representing ten of nineteen publically funded Ontario paramedic training programs. All were at the entry to practice level (i.e., all seeking their first authorization to perform delegated acts), all provisionally employed and all already certified by the Ministry of Health Emergency Health Services Branch in Ontario. Mean scores by station and dimension, along with cut scores and fail rates by dimension are provided in Table 2.

Evidence for scoring

Exploring scoring details allowed us to identify whether there were any problematic redundancies or items in the rating tool and in rating behaviors. First, we explored whether dimensions on the rating tool would demonstrate evidence of independence. Inter-item and item-total correlations for the GRS across all stations ranged from $r = .46$ (HG and PS; RU and PS) to $.65$ (DM and SA), suggesting that dimensions were indeed functioning and/or treated independently with no obvious signs of redundancy (see Table 3 for individual results) (Streiner and Norman 2008).

Second, our generalizability results confirmed very low error variance attributable to both items and raters (nested in stations); 1.8 and .13% respectively. Third, our results revealed an inter-rater reliability of .91 when averaged across seven stations. When calculated by dimension only, inter-rater reliability ranged from .76 (patient assessment) to

Table 2 Mean scores combining all candidates (reported as percentage) by station and performance dimension, cut score calculated using a borderline group method by dimension and the proportion of candidates who scored below the cut score by dimension

Dimension	ST-1	ST-2	ST-3	ST-4	ST-5	ST-6	ST-7	Mean	Cut score	Fail rate
SA	72.9	75.5	76.4	70.4	70.7	69.2	72.7	72.5	57.4(1.2)	2.9
HG	72.1	71.5	76.3	73.6	73.2	72.7	72.9	73.2	64(1.0)	8.5
PA	69.0	71.5	77.7	71.5	70.0	69.5	72.6	71.7	57.4(1.2)	1.4
DM	66.7	70.5	71.7	69.6	67.6	68.1	71.3	69.4	50.0(1.0)	.0
RU	74.6	73.3	76.5	75.9	71.4	76.5	74.0	74.6	65.6(1.1)	5.7
CM	76.0	75.6	78.6	78.0	74.0	75.4	74.1	75.9	65.6(1.1)	7.1
PS	70.8	70.0	73.4	69.3	71.4	70.9	73.2	71.3	57.1(1.2)	4.3
Mean	71.7	72.6	75.8	72.6	71.2	71.7	72.9			

SA situation awareness, HG history gathering, PA patient assessment, DM decision making, RU resource utilization, CM communication, PS procedural skill, ST station

Table 3 Inter-item and item total correlations between dimensions included on the global rating scale

Effect	Inter-item correlations						Item-total correlations
	SA	HG	PA	DM	RU	CM	
Situation awareness (SA)							.72
History gathering (HG)	.56						.70
Patient assessment (PA)	.55	.56					.69
Decision making (DM)	.65	.50	.58				.73
Resource utilization (RU)	.58	.63	.51	.54			.70
Communication (CM)	.54	.62	.51	.53	.64		.69
Procedural skill (PS)	.53	.46	.55	.63	.46	.47	.65

.82 (procedural skill), also averaged across seven stations. See Table 4 for facet variance components, percentage of variance attributable to each, and reliability results.

Evidence for generalization

We were interested in exploring how well the observed sample represents the broader range of possible performances. Our sampling strategy involved creating curriculum maps using guiding practice documents (described above). Across seven stations, 54% (7/13) of the relevant medical directives were included. Two of the stations did not include any medical directives as a way of further reflecting practice but in a different way (e.g., having to rule out select care options). Of the skills deemed relevant to support the medical directives, 82% ($n = 14/17$) were included. All three pre-defined age groups (adult $n = 5$ -stations; pediatric $n = 2$ -stations; geriatric $n = 2$ -stations) were represented, and of 20 broadly classified disease types deemed priorities (e.g., cardiovascular, respiratory, endocrine) 50% ($n = 10$) were included. Lastly, of 69 relevant (i.e., practice based) national competencies, 84% ($n = 58$) were represented across the seven stations. See supplemental material sections 1–4 for additional details.

In regards to our hypothesis that scores obtained in this setting would be similar if an entirely new set of cases were used, our results revealed a generalizability coefficient (reliability) of .70 (inter-station = .77). The “candidate \times station” facet (i.e., context specificity) resulted in the second highest source of error variance next to random (unidentifiable) error (see Table 4 for individual variance components and percentage of error attributable to each facet). A decision study, which allows for predictions of generalizability/reliability assuming modifications are made in future exams (e.g., stations, rater, items) suggests approximately 11 stations may be needed to reach a reliability of .80, assuming all other facets (e.g., candidate variance) remained stable.

Evidence for extrapolation

We explored whether performance in an assessment context is associated with performance in the real world. Of the 125 who completed the exam, we obtained clinical data for 107 candidates (after removing those who were unsuccessful and for whom we did not have complete data). The total number of patient contacts was 24,880 or an average of 232.5

Table 4 Generalizability theory analysis results. Inter-station and inter-rater reliability results including individual variance components and percentage of total variance

Facet	Variance component	% of total var.
Candidate	.171	11.26
Station	.017	1.12
Rater: station	.002	.13
Items	.027	1.78
Candidate × station	.303	19.96
Candidate × rater: station	.216	14.23
Candidate × item	.008	.53
Station × item	.008	.53
Rater × item: station	.005	.33
Candidate × station × item	.136	8.96
Candidate × rater × item: station	.625	41.17
	1.52	100
Reliability		
Overall: $G1 = .11/G7 = .70$		
$V(c)/[V(c)] + [V(s)/7] + [V(r : s)/14] + [V(cs)/7] + [V(cr : s)/14]$ $+ [V(ci)/7] + [V(si)/7] + [V(ri : s)/14] + [V(csi)/7 * 7] + [V(cri : s)/14 * 7]$		
Inter-rater reliability: $G1 = .42/G7 = .91$		
$V(c) + [V(cs)/7] + [V(ci)/7] + [V(csi)/7 * 7]/[V(c)] + [V(cs)/7] + [V(ci)/7]$ $+ [V(csi)/7 * 7] + [V(r : s)/14] + [V(cr : s)/14] + [V(ri : s)/14] + [V(cri : s)/14 * 7]$		
Inter-station reliability = $G1 = .26$; $G7 = .77$		
$V(c) + [V(cs)/7] + [V(ci)/7] + [V(csi)/7 * 7]/V(c) + [V(cs)/7] + [V(ci)/7]$ $+ [V(csi)/7 * 7] + [V(r : s)/2] + [V(cr : s)/2] + [V(ri : s)/2] + [V(cri : s)/2 * 7]$		

(SD=76.5) per candidate over 210 days. The total number of errors was 198, with an overall mean rate of error of 1.85 (SD = 1.6) per person over 210 days. After taking into account call volume and patient contacts the mean proportion of errors per paramedic was .76% (SD = 0.78) of patient contacts.

We explored whether participants' simulation-based assessment scores on each of the seven performance dimensions across seven stations were associated with their committing clinical errors. We found low, non-significant correlations between the performance dimensions and either the total number of errors (from $r = -.12$ to $-.01$), or the proportion of errors (from $r = -.05$ to $.00$).

Evidence for implications

We explored how the decision about performance, such as pass or fail impact the profession and/or society (i.e., implications) by looking for clinical errors in practice. Table 2 contains data on mean station and dimension scores, cut scores by dimension and the "fail rate" by dimension. Eight of 125 candidates (6.4%) did not achieve the cut score on one or more dimensions. We compared the total number of errors between these new data and data from previous performance based assessment models (as described above) and found

no significant differences [NEW mean = 1.8 (SD = 1.6), 95% CI = 1.5–2.1 vs. PREVIOUS = 2.1 (SD = 1.7), 95% CI = 1.8–2.4; $F(1) = 1.70$, $p = .19$]. When we conducted the same analyses taking into consideration number of patient interactions (as opposed to just time) we also found no significant difference [NEW = .78 (SD.72), 95% CI = .64 to .92 vs. PREVIOUS = .74 (SD.84), 95% CI = .58 to .89; $F(1) = .17$, $p = .68$].

Discussion

In this study we examined validity evidence associated with a paramedic simulation based assessment using Kane's validity framework, one of many validity frameworks. Specifically, we evaluated whether the assessment scores generated in the performance test could be used to inform decisions regarding candidates' readiness for entry to paramedic practice and to predict their future clinical performance. As suggested by Kane, the final step requires synthesizing the evidence to formulate a validity argument, which helps to clarify whether additional evidence is needed to strengthen future interpretation-use arguments (IUAs). (Kane 2012) Our synthesis suggests the data supported our claims related to the scoring, generalization and implication inferences, however, the data do not support our claims related to the extrapolation inference. Our research processes and outcomes have specific implications for the paramedicine context, and as a worked example, also provide general principles, lessons and decision points to consider when applying Kane's validity framework to simulation-based assessments.

To develop a validity argument, Kane suggests evaluating the evidence and deciding whether to accept the IUA, reject it, and/or revise the process and/or proposed use. (Kane 2012; Cook et al. 2015) In our case, we cannot change the proposed use because the assessment group continues to need a high-stakes assessment of clinical competence. Instead, changes to the assessment process or assessment tool, outcomes measures and/or gathering of further evidence are more appropriate steps. Our perspective is that no one study fully "validates" or "invalidates" an assessment strategy, rather validity is a matter of degree with more or less evidence in support of certain claims or inferences. We interpret our finding for the extrapolation inference below (which of the four inferences we examined, refutes out IUA), and propose changes to the IUA and assessment process accordingly.

To date, evidence collected for the extrapolation inference of our performance assessment suggests candidates' scores relate positively to scores generated using direct observation of their behaviour in the workplace using the same global rating scale. (Tavares et al. 2012, 2014) We attempted to relate candidates' performance scores to a different workplace outcome, their clinical error rates, which was not successful. Interpreting this result may mean that the performance assessment failed to screen candidates sufficiently enough to predict their eventual errors, or, alternatively, that detection of clinical error rates may not be sensitive enough to clinically relevant differences in performance ability or competence. Further, perhaps the performance assessment only passed candidates performing at a high level, resulting in limited variation in clinical practice and thus a limited capacity to predict strong and poor performers in practice. As with our original assumptions, these interpretations can now be evaluated to generate additional extrapolation validity evidence. As such, we suggest additional studies focus on the extrapolation inference, either by studying how the performance assessment relates to other

clinical measures, by seeking additional evidence relating use of the global rating scale in both the simulated and workplace settings or by seeking more sensitive measures or ability in practice.

Relating our present findings to the existing body of evidence, we argue that our performance assessment appears to be screening candidates appropriately. That is, while the evidence supports our scoring, generalization, and implications inferences, we suggest the evidence for the “extrapolation” inference only suggests that we have less favourable evidence than what we set out to achieve. Even if all tests in this study supported our IUA, we would still recommend testing additional assumptions using qualitative or quantitative strategies to pursue further sources of evidence. For example, the low candidate variance identified in our G-study and the consistent means across stations suggests we assessed a highly homogenous group of candidates, or that some improvements could be made to discriminate between candidates at the tool, rater or station level. We note these issues, despite our relatively positive findings, to illustrate that our validity argument is not a conclusion, but instead represents a positive step in a series of studies aimed at establishing and refining the validity evidence for our OSCE.

Applying Kane’s validity framework involved translating a philosophy and theoretical model into practice-oriented steps, interpretations and judgments. Reflecting on our experience, the framework helped us structure and organize our thinking. Specifically, we felt the analogy of viewing each inference as links in a chain or segments of a bridge helped us define the inter-relationships between our hypotheses, analyses, and interpretations. Further, beginning with the IUA helped us focus our efforts. By contrast, we experienced challenges when deciding how to prioritize collecting and reporting of evidence for the four inferences. We recognized that the weakest and most questionable assumptions, as they related to our IUA, must be prioritized in the study design. We also recognized, based on Kane’s writing (Kane 2012), that targeting some assumptions, before there is sufficient evidence for others, could leave important gaps in the validity argument. Knowing these principles, we still grappled with the subjectivity of prioritizing the “weakest and most questionable” assumptions. In the end, we decided to consider our IUA as the priority, rather than treating the inferences like a sequential checklist. Researchers might be helped to make their decisions using Cronbach’s criteria including: (a) prior uncertainty, (b) information yield, (c) cost, and (d) leverage (what matters to end-users). (Cronbach 1989) Ultimately, we learned that researchers make a number of judgments during validation, and thus must provide strong rationales for their decisions, especially where conventions or firm benchmarks are unavailable.

Limitations

There are some limitations to consider. First, the data sources used to calculate clinical errors in practice, while robust, currently lack reliability data, meaning we could not calculate degree of attenuation associated with those data. Also, the data depend on accurate paramedic documentation and we are unable to confirm the level of accuracy reported. Second, some of our analyses may have been affected by our sample size, both in the number of candidates considered in this analysis and in the length of time we monitored paramedics in practice. Third, this cohort of paramedics represents a very homogenous and highly selected group, which likely affects our results. Other settings with different ranges in ability may yield different results. Finally, our extrapolation and implications evidence was based on only candidates who successfully completed the certification process, resulting in low proportion of errors. This may tell us that the screening process is working

well but also attenuate our results by this methodological limitation (i.e., providing individuals who scored below the standard access to practice, would be unethical).

Conclusions

Assessing clinical competence is complex, with numerous assumptions threatening validity claims. Ensuring the trustworthiness and defensibility of assessment decisions requires that these assumptions be identified, tested, and related to the intended interpretations and uses of the assessment scores. We applied Kane's validity framework and conclude that we must accept a revised version of our original interpretation/use argument: this seven station simulation-based exam can be used to assess clinical competence at the entry to practice level for paramedics, though it cannot be used to predict future clinical error rates, though it may well be that error rates are an inappropriate outcome as measured in this study. Our experiences highlight how judgment is an important part of validation strategies at both the individual inference level, and after all evidence has been generated. Our report meets other researchers' calls for studies that use formal validation frameworks (Cook and Hatala 2016; Brennan 2013), and our modeling of the process is aimed to stimulate additional studies from the simulation community and beyond.

Acknowledgements The authors would like to thank the Ontario Base Hospital Group for their support in completing this study.

References

- Brennan, B. L. (2001). *Generalizability theory*. New York, NY: Springer.
- Brennan, B. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement*, 50(1), 74–83.
- Clauser, B. E., Margolis, M. J., Holtman, M. C., Katsufakis, P. J., & Hawkins, R. E. (2012). Validity considerations in the assessment of professionalism. *Advances in Health Sciences Education, Theory, and Practice*, 17(2), 165–181.
- Cook, D., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560–575.
- Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Health Sciences Education Theory and Practice*, 1(1), 31.
- Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education Theory and Practice*, 19(2), 233–250.
- Cronbach, L. J. (1989). Construct validation after thirty years. *Intelligence Measurement Theory and Public Policy*, 3, 147–171.
- Frank, J., Snell, L., & Sherbino, J. (2014). *Draft Can-MEDS 2015 physician competency framework-series III*. Ottawa, Ontario: The Royal College of Physicians and Surgeons of Canada.
- Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the objective structured assessment of technical skills (OSATS): A systematic review of validity evidence. *Advances in Health Sciences Education Theory and Practice*, 20(5), 1149–1175.
- Humphrey-Murto, S., & MacFadyen, J. (2002). Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Academic Medicine*, 77(7), 729–732.
- Kane, M. T. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 20(10), 5–17.
- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17.
- Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

- Kane, M. T. (2013b). Validity. In B. L. Brennan (Ed.), *Educational measurement*. Westport, CT: Praeger Publishers.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*. McMillan: Old Tappan, NJ.
- Ministry of Health and Long-Term Care, Emergency Health Services Branch. (2007). *Basic life support patient care standards version 2.0*. Toronto, Ontario: Publications Ontario.
- Ministry of Health and Long-Term Care, Emergency Health Services Branch. (2015). *Advanced life support patient care standards version 3.2*. Toronto, Ontario: Publications Ontario.
- Mylopoulos, M., & Regehr, G. (2011). Putting the expert together again. *Medical Education*, 45(9), 920–926.
- Paramedic Association of Canada. (2016). National occupational competency profile 2016. Retrieved February 22, 2017, from <http://paramedic.ca/site/nocp?nav=02>.
- Ponton-Carass, J., Kortbeek, J. B., & Ma, I. W. Y. (2016). Assessment of technical and nontechnical skills in surgical residents. *The American Journal of Surgery*, 212(5), 1011–1019.
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczyńska, U. (2011). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85(2), 370–395.
- St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education Theory and Practice*. <https://doi.org/10.1007/s10459-016-9716-3>.
- Streiner, D. L., & Norman, G. R. (2008). *Health Measurement Scales: A practical guide to their development and use* (4th ed.). Oxford, New York: Oxford University Press.
- Sunnybrook Centre for Prehospital Medicine. (2016). *Regional Base Hospital 2015–2016 annual report*. Toronto: Ontario.
- Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K. (2012). Global rating scale for the assessment of paramedic clinical competence. *Prehospital Emergency Care*, 17(1), 57–67.
- Tavares, W., Bowles, R., & Donelon, B. (2016). Informing a Canadian paramedic profile: Framing concepts, roles and crosscutting themes. *BMC Health Services Research*, 16, 477.
- Tavares, W., LeBlanc, V. R., Mausz, J., Sun, V., & Eva, K. W. (2014). Simulation-based assessment of paramedics and performance in real clinical contexts. *Prehospital Emergency Care*, 18(1), 116–122.
- Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189–205.