Global Rating Scale for the Assessment of Paramedic Clinical Competence

Walter Tavares, ACP, PhD, Sylvain Boet, MD, Med, Rob Theriault, CCP, BHSc, Tony Mallette, ACP, Kevin W. Eva, PhD

Abstract

Objective. The aim of this study was to develop and critically appraise a global rating scale (GRS) for the assessment of individual paramedic clinical competence at the entry-to-practice level. **Methods.** The development phase of this study involved task analysis by experts, contributions from a focus group, and a modified Delphi process using a national expert panel to establish evidence of content validity. The critical appraisal phase had two raters apply the GRS, developed in the first phase, to a series of sample performances from three groups: novice paramedic students (group 1), paramedic students at the entry-to-practice level (group 2), and experienced paramedics (group 3). Using data from this process, we examined the tool's reliability within each group and tested the discriminative validity hypothesis that higher scores would be associated with higher levels of training and experience. **Results.** The development phase resulted in a seven-dimension, seven-point adjectival GRS. The two independent blinded raters scored 81 recorded sample performances ($n = 25$ in group 1, $n = 33$ in group 2, $n = 23$ in group 3) using the GRS. For groups 1, 2, and 3, respectively, interrater reliability reached 0.75, 0.88, and 0.94. Intrarater reliability reached 0.94 and the internal consistency ranged from 0.53 to 0.89. Rater differences contributed 0–5.7% of the total variance. The GRS scores assigned to each group increased with level of experience, both using the overall rating (means = 2.3, 4.1, 5.0; $p < 0.001$) and considering each dimension separately. Applying a modified borderline group method, 54.9% of group 1, 13.4% of group 2, and 2.9% of group 3 were below the cut score. **Conclusion.** The results of this study provide evidence that the scores generated using this scale can be valid for the purpose of making decisions regarding paramedic clinical competence. **Key words:** educational measurement; clinical competence; licensure; certification; paramedics; global rating scale; rating scales

PREHOSPITAL EMERGENCY CARE 2012;Early Online:1–11

Received March 9, 2012, from the Centennial College Paramedic Program (WT), Toronto, Ontario, Canada; the University of Toronto, Wilson Centre for Medical Education (WT), Toronto, Ontario, Canada; the Department of Anesthesiology, the Ottawa Hospital (SB), the University of Ottawa Skills and Simulation Centre (SB), and the Academy for Innovation in Medical Education (SB), University of Ottawa, Ottawa, Ontario, Canada; Georgian College Paramedic Program (RT), Barrie, Ontario, Canada; Lambton College Paramedic Program (TM), Sarnia, Ontario, Canada; and the Centre for Health Education Scholarship, Department of Medicine, University of British Columbia (KWE), Vancouver, British Columbia, Canada. Revision received April 27, 2012; accepted for publication May 16, 2012.

Supported by the Applied Research and Innovation Centre at Centennial College in Toronto, ON, Canada.

The authors would like to thank all the students, paramedics, and experts who agreed to participate in this study. The authors would also like to thank (from east to west) the Emergency Health Service of Nova Scotia; Centennial College Science and Technology Centre/University of Toronto, ORNGE, Sunnybrook Centre for Prehospital Medicine, Georgian College, Lambton College of Applied Arts and Technology, and York Region Emergency Medical Services of Ontario; Red River College of Manitoba; Saskatchewan Institute of Applied Science and Technology; Lakeland College, University of Alberta and Professional Medical Associates of Alberta; Justice Institute of British Columbia; and the Society for Prehospital Educators of Canada (SPEC) for their contributions.

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Address correspondence and reprint requests to: Walter Tavares, School of Community and Health Studies, Centennial College, P.O. Box 631 Station A, Toronto, ON, Canada, M1K 5E9. E-mail: wtavares@centennialcollege.ca

doi: 10.3109/10903127.2012.702194

Introduction

Paramedics provide emergency and nonemergency care to patients suffering from diverse complex medical conditions and traumatic injuries. The level of clinical competence expected of paramedics has grown and, like other health professions, lack of competence can adversely affect patient safety and outcomes. Educational institutions, employers, licensing bodies, and/or regulators have a responsibility to ensure that paramedic candidates entering the profession are ready for independent practice. Performance-based examinations are an integral part of ensuring clinical competence.[1] Administering them well requires evidence of adequate reliability and validity.[2]

In the field of paramedicine, assessment tools have generally taken the form of task-specific checklists.[3–5] For example, the National Registry of Emergency Medical Technicians includes as part of its examination a series of task-specific (often chronological) binary checklists describing performance expectations along with critical errors for each task (http://www.nremt.org). While checklists may be appropriate in some contexts (e.g., isolated procedural tasks), paramedic clinical competence, which includes both technical and nontechnical elements as well as variations in process, may be challenging to identify

and measure using checklists.[6,7] Global rating scales (GRSs) have rarely been considered for the assessment of paramedic competence,[8,9] despite their apparent necessity and associated advantages.[7,10,11] Global rating scales are subjective, but they have been demonstrated to better differentiate levels of experience when compared with checklists,[7,12,13] although checklists may better differentiate between individuals within a novice group of examinees.[14] Recently, Martin et al. revealed considerable variability in the rate at which raters reported errors when observing videotaped performances of paramedic practice.[9] In general, however, there is a paucity of research evaluating the reliability and validity of assessment tools in paramedicine, and, therefore, no gold standard (checklist or GRS) exists.

Kane (2006) describes validation as a process of evaluating proposed interpretations of data based on the scores generated from an instrument. This involves clearly stating intended interpretations, identifying assumptions, and critically evaluating the assumptions associated with the measurement tool.[15] The lack of this type of research in paramedic settings raises concerns regarding the accuracy and defensibility of the performance-based assessments used. This study aimed to develop and critically appraise a generic global rating scale to measure individual paramedic clinical competence for summative "entry-to-practice" decisions.

METHODS

Ethical approval was provided by Centennial College in Toronto, Ontario, Canada (approval # 087) and informed consent was obtained from each participant. We structured our study in two phases: a *development phase,* in which we ensured that the construct of paramedic clinical competence was adequately represented, and an *appraisal phase,* in which we critically evaluated intended interpretations using the GRS.

Study Design for the Development Phase

The development phase of this study focused on ensuring that the target construct of primary care paramedic clinical competence was adequately represented by the rating scale. Local experts, a broad focus group, and then a national panel of experts representing a variety of stakeholders in paramedic education, certification, and employment engaged in an iterative process until there was evidence of agreement that the construct was adequately represented by the scale.

First, local experts (WT and two other faculty from Centennial College) engaged in task analysis aimed at identifying relevant behaviors in paramedic clinical practice through observation of various clinical cases completed both in simulation and in actual clinical practice. This group then clustered the behaviors observed and ultimately identified specific dimensions so that a working prototype GRS could be prepared. Next, a focus group representing end users (i.e., educators, certifying bodies) and raters was assembled to evaluate the characteristics, items, definitions, and language used in the working prototype. This focus group was given an opportunity to apply the working prototype to a series of sample video performances to identify concerns, issues, or gaps. Finally, a national expert panel engaged in a modified Delphi process to complete the development phase.[16,17] This involved presenting experts with a sample list of dimensions with bulleted statements intended to define each (developed and refined in task analysis and focus-group session), using a Web-based survey tool. Experts were asked to rate the relevance of each element (i.e., dimension and descriptor, seven-point labels and descriptors) as it relates to the intended construct using a four-point scale from 1 = not relevant to 4 = extremely relevant, rate their level of agreement with regard to adequate representation of the construct using a four-point scale from 1 = strongly disagree to 4 = strongly agree, and provide comments. Results were then shared with each expert panel independently in subsequent rounds.

Analysis

Local experts continued the task analysis until saturation (i.e., until no new distinct and relevant behaviors were identified). Clustering continued iteratively until all behaviors were organized, dimensions could be identified, and a working prototype GRS was prepared. The focus group engaged in open discussion facilitated by the principal investigator (WT) until saturation (i.e., no new changes/revisions were suggested). Finally, using the working prototype GRS, the modified Delphi process continued until consensus was reached (defined as 80% agreement) among national experts on all items, definitions, rating labels, and rating label definitions.

Study Design for the Critical Appraisal Phase

The critical appraisal phase of this study tested the following hypotheses: 1) that the dimensions listed in the prototype are distinct and adequately represent the construct of interest, 2) that individual paramedics can be consistently differentiated by raters using the GRS (i.e., that the tool is reliable), and 3) that higher scores are empirically associated with higher levels of experience when using the GRS to rate paramedic clinical performance. To test these hypotheses, we subjected the scale to a quasi-experimental design to evaluate internal structure, reliability, and relationship to other variables. This involved first recording clinical performances by three distinct groups—*novice* (in training) paramedic students (group 1), *entry-to-practice* (about-to-graduate) paramedic students (group 2), and *experienced* paramedics (group 3)—all of whom

completed the same case in simulation. The videos were coded, randomized, and distributed to raters to score using the prototype GRS.

Participants and Scenario

Purposive sampling was used to recruit paramedic students for groups 1 (novice) and 2 (entry-to-practice) from a local paramedic program and to recruit participants for group 3 (experienced paramedics) from six different emergency medical services in Southern Ontario, Canada. Our selection of these groups was based on evidence suggesting that expertise develops as a result of a greater knowledge base,[18] greater experience through supervised and unsupervised exposure to a variety of patients,[18,19] and more opportunity for deliberate practice.[19] This provided a range of competence that could then be used to test the scale's ability to differentiate between levels of performance.

Participants in each group were required to complete the same case in a high-fidelity simulator. SimMan (Laerdal Medical, Stavenger, Norway) was placed in a mock ambulance equipped with audio and video recording equipment. The simulated scenario involved an unstable cardiac patient with decreased level of consciousness who, at a predetermined marker, deteriorated to cardiac arrest over 9 minutes. Research assistants playing the role of transfer company were present as part of the simulation to provide a history to the examinee by answering his or her queries. Because the scenario was set at the "side of a roadway," as described in Appendix 1 (available online), the participants also had to demonstrate awareness of surveying the scene and ensuring the safety of those involved. The scenario used was similar to those included in traditional entry-to-practice assessment processes and required a broad range of technical and nontechnical skills. It was based on an actual clinical case, piloted, and refined using paramedic education and simulation experts, students, and active paramedics. Provincial and national scope-of-practice guidelines informed the case-development process.[20–23] The participants were instructed to assess and manage the clinical case to the best of their ability using any available equipment and resources. Content and performance expectations were carefully considered to ensure that all groups (including group 1) had sufficient knowledge and skill to complete the case to standard without identifying themselves as being at a particular level of training. We intentionally selected paramedics for group 3 with less than five years of experience to reduce heterogeneity and limit differences in appearance that could bias the ratings.

Sample Size Calculation

The primary outcome for this study was the ability of raters to differentiate between groups 1, 2, and 3 using the prototype GRS. With an estimated effect size of 0.8, which is widely accepted in education and psychology to indicate a large effect[24] and has been used in similar scale validation studies,[25] a two-tailed α of 0.05, and a β of 0.20, 25 participants were required per group, for a total of 75 participants. Rounding up to account for potential attrition, we sought to enroll 85 participants. This sample size facilitated secondary outcome (e.g., item analysis, interrater reliability) analyses as well.[26]

Rating Procedure

Videos were given a study code, randomly ordered to minimize potential confounders, and then distributed to two blinded independent raters (described below) who were asked to score the videos as they normally would for entry-to-practice decisions. The raters were allowed to take notes while observing the performance, but rewinding or pausing the video was not allowed, in order to replicate natural conditions as closely as possible. The raters were informed that the videos represented a collection of performances from a variety of clinicians, and that each was to be evaluated independently using the GRS. Prior to scoring, the raters were provided with a brief introduction to the rating scale and given instructions on how to apply the scale. One sample case from pilot recordings was used to allow the raters to practice applying the scale before beginning data collection. Both raters and the principal investigator (WT) met following the initial practice rating to discuss any rating issues. The introduction, instructions, practice sessions, and discussion took approximately 60 minutes. We intentionally limited rater training to evaluate outcomes under the most natural conditions. No attempt was made to calibrate the raters, and only scale application issues were discussed. Two months following the initial rating of all videos, each rater was randomly assigned a subset of the videos to enable an evaluation of intrarater reliability.

Analysis

To test assumptions related to the scale's content, internal structure was analyzed via item analysis (i.e., internal consistency, interitem and item–total correlations).[27] Variance attributable to raters, items, and the relevant interactions between those facets was determined using generalizability theory and used to calculate reliability, inter and intra-rater reliability. Inter- and intrarater reliability was calculated using generalizability theory.[28] A modified borderline group method[29–31] for establishing cut scores was also applied and reported using descriptive statistics. This involved having raters judge candidate performance using a seven-point adjectival scale that was included at the end of the GRS. They were asked to rate each

candidate's overall performance as either unsatisfactory (1 = unsafe, 2 = unsatisfactory, 3 = poor/weak) or satisfactory (4 = marginal, 5 = competent, 6 = highly competent, 7 = exceptional). Prior to the rating task, the raters were informed that scores of 3 or 4 on this overall category would be considered the borderline group. Scores assigned on the seven construct-specific dimensions for that cohort of candidates were aggregated to establish a cut score by dimension. Finally, using scores from each of the three groups, we tested the hypothesis that higher scores are related to higher levels of experience using analysis of variance (ANOVA). All data were analyzed using SPSS Version 19 (IBM, Armonk, NY) and generalizability software (G String Version IV, Bloch R, Hamilton, Ontario, Canada). The level of significance was set at $p = 0.05$ (two-tailed).

RESULTS

Development Phase

Task Analysis and Item Development

The development phase involved having experts from a paramedic program conduct a task analysis[32] using multiple simulation-based paramedic clinical performances and actual clinical cases. Experts identified 257 observable behaviors from a variety of contexts and then iteratively arranged the behaviors into clusters relevant to paramedic practice. Additional performance observations continued to determine sufficiency of the list and/or the need for further refinements to the clusters.

Eight dimensions in total were identified: situation awareness, history gathering, patient assessment, decision making, implementation, resource utilization, communication, and procedural skills. Using the behaviors identified during the task analysis, descriptors with examples for each were attached to each dimension. An initial working prototype GRS including the eight dimensions and seven-point adjectival scales was created. A seven-point scale was selected to facilitate reliability without creating levels between which the raters would have difficulty differentiating.[26,33] Rating labels, with definitions for each (based on practice standards, patient safety, and readiness for independent practice or progression) anchored each of the seven points.

Focus Group

Next, a focus group of 17 practicing paramedic clinicians who were also practicing educators and assessors from five different emergency medical services in southern Ontario, Canada, contributed to the refinement of the scale. The raters reviewed and approved the list of dimensions assembled, the definitions associated with each dimension, the rating labels selected, and their definitions. After having an opportunity to apply the scale to two prerecorded videos of paramedic simulations, the dimension "implementation" was identified as a source of disagreement regarding its distinction from other dimensions. The focus group along with the researchers elected to retain the dimension for the national expert panel.

National Expert Panel

Nine experts from five provinces across Canada participated in a modified Delphi process.[16,17] The experts were selected based on their individual experience in paramedicine and unique perspective relative to the rating scale's intended application (i.e., entry-to-practice decisions). For example, some were responsible for graduating paramedic candidates ($n = 5$), while others were responsible for employment ($n = 2$) or for certification ($n = 2$). All were practicing experienced educators ($n = 9$, median of 10 years in paramedic education), researchers ($n = 3$), or active paramedics ($n = 8$, median of 15 years in clinical practice).

Round 1 of the Delphi process achieved consensus (>80% agreement) on all dimensions, rating labels, and definitions except for the dimension "implementation." Following round 1, bulleted statements were converted to general descriptions for each dimension and suggestions for revisions were implemented or shared with the group for consensus prior to round 2. In round 2, a revised GRS was distributed and achieved consensus on all levels, with the exception of the "implementation" dimension. Similar to the focus-group session, the expert panel disagreed regarding its distinction from other dimensions and its inherent inclusion in each. Based on the feedback from the focus group, and results of the Delphi process, the dimension "implementation" was eliminated from the rating scale and the Delphi process discontinued. A copy of the final GRS is included in Appendix 2 (available online).

Critical Appraisal Phase

Group Participants and Raters

Participants for each group were enrolled between January and May 2011. Eighty-five participants were enrolled. Twenty-five novice paramedic students (17 men, 8 women) in group 1, 36 entry-to-practice students (19 men, 17 women) in group 2, and 24 active paramedics (14 men, 10 women) in group 3. Three videos in group 2 were discarded because of technical difficulties. One video in group 3 was discarded once a participant disclosed he was not an active paramedic. Of the remaining 23 paramedics in group 3, the mean years of experience was 2.4. This group represented six different paramedic services and six different paramedic programs in Ontario, Canada. A total of 81 videos, each lasting 9 minutes, were

TABLE 1. Variance Components and Percentage of Total Variance by Group

Effect	G1 VC	% of Total Variance	G2 VC	% of Total Variance	G3 VC	% of Total Variance
Person	0.89	41.1%	1.05	47.8%	0.76	32.2%
Rater	0.00	0.1%	0.00	0.0%	0.14	5.7%
Item	0.07	3.2%	0.00	0.0%	0.06	2.4%
Person × rater	0.11	5.1%	0.24	11.1%	0.34	14.5%
Person × item	0.25	11.5%	0.17	7.9%	0.10	4.2%
Rater × item	0.24	11.3%	0.23	10.3%	0.19	8.1%
Person × rater × item	0.59	27.1%	0.50	22.9%	0.77	32.78%
TOTAL VARIANCE	2.16	100%	2.19	100%	2.35	100%

G1 = group 1: novice-level paramedic students; G2 = group 2: entry-to-practice-level paramedic students; G3 = group 3: experienced active paramedics; VC = variance components.

submitted to two raters for scoring. The initial rating procedure was completed over a one-month period. The subsequent rating procedure (used to calculate intrarater reliability) involved rating a random selection of 30 of the 81 videos. This second rating took place two months following the initial rating task and was completed over a two-week period.

Two raters were selected from two different paramedic programs in Ontario, Canada. Together the raters averaged 11 years' experience as paramedic educators, 22 years as paramedics, and 13 years evaluating clinical performances. All videos were scored on seven dimensions and given an "overall" performance rating by each rater.

Reliability

Reliability analyses were conducted on each group independently to avoid artificially inflating the heterogeneity in the videos. The proportion of variance attributable to rater differences in groups 1, 2, and 3, respectively, was 0.1%, 0%, and 5.7% of group total variance. All variance components are illustrated in Table 1.

Using participant as the facet of differentiation and items as the facet of generalization, internal consistency was calculated and found to be 0.89, 0.71, and 0.53 for groups 1, 2, and 3, respectively. Inter-item correlations ranged from 0.62 to 0.93 and item–total correlations ranged from 0.74 to 0.92. Individual inter-item and item–total correlations are provided in Table 2 along with the correlation between each item and the overall rating assigned.

The interrater reliability for groups 1, 2, and 3 reached 0.75, 0.88, and 0.94, respectively. Intrarater reliability was calculated using the scores assigned to the 30 randomly selected videos and reached 0.94. The reliability for each dimension considered independently ranged from 0.54 (communication) to 0.83 (decision making) within group 2 (i.e., those selected from the target population). Individual generalizability coefficients (G coefficients) for each dimension are reported in Table 3.

Relationship to Other Variables

To test for evidence of discriminative validity, using all dimensions, a one-way ANOVA was performed using group as the independent variable and average score as the dependent variable. The effect of group was found to be statistically significant both based on overall scores ($F(2,78) = 29.5$, $p < 0.001$) and for each individual dimension (see Table 4). Making pairwise comparisons, the differences between the means across group aligned with expectations in 23 out of 24 instances, consistency that can be expected to occur less than 0.1% of the time according to binomial probability theorem (i.e., $p < 0.001$). The one reversal (group 2 > group 3 in the "Communication" skills dimension) was slight, with an observed difference of 0.06.

We applied the modified borderline group method to each dimension to evaluate the relationship between failure rate and group assignment and found the highest failure rates in group 1, followed by group 2, and then group 3. The results are provided in Table 5.

TABLE 2. Interitem and Item–Total Correlations Using Data from All Three Groups

Dimension	SA	HG	PA	DM	RU	COM	Item–Total Correlation	Correlation with "Overall" Rating
Situation Awareness (SA)							0.93	0.95
History Gathering (HG)	0.71						0.74	0.74
Patient Assessment (PA)	0.93	0.69					0.89	0.91
Decision Making (DM)	0.92	0.70	0.88				0.92	0.95
Resource Utilization (RU)	0.81	0.67	0.77	0.79			0.85	0.84
Communication (COM)	0.69	0.63	0.62	0.66	0.76		0.74	0.75
Procedural Skill (PS)	0.85	0.66	0.81	0.89	0.75	0.67	0.88	0.92

TABLE 3. Interrater Reliability for Each Dimension Calculated Using Generalizability Theory and Group 2 Data

Dimension	G Coefficient
Situation Awareness	0.83
History Gathering	0.64
Patient Assessment	0.81
Decision Making	0.84
Resource Utilization	0.60
Communication	0.54
Procedural Skill	0.67

G coefficient = generalizability coefficient.

DISCUSSION

Kane (2006) describes validation as a process of generating an *interpretive argument* in which proposed interpretations are clearly stated (for example, higher scores on the GRS are indicative of a higher level of experience) and then critically evaluated for plausibility and coherence.[15] Numerous assumptions between the observation of performance and the final decision regarding competence must be identified and evaluated if the interpretation based on scores generated is to be considered defensible. Using this framework, we proposed that the scale would be used to make inferences regarding paramedic clinical competence at the entry-to-practice level. We then identified construct representation as the first assumption to be tested and hypothesized that higher scores would align with higher levels of experience. The results of a development and critical appraisal process included in this study suggest that this GRS can be implemented in a way that provides reasonable reliability and capacity to differentiate both between groups and between individuals within group and, therefore, enables inferences regarding paramedic clinical competence.

The development phase and early stages of the critical appraisal phase of this study were aimed at evaluating the adequacy of construct representation. This content validation process involved clinicians, educators, and experts in the field of paramedicine collectively and iteratively ensuring an appropriate focus of the GRS. This involved detailed task analyses (i.e., observation of clinical performance using a wide variety of cases) in simulation and clinical settings, a large focus group of clinicians who were also educators and raters, and a national expert panel representing a variety of stakeholders responsible for making decisions regarding independent paramedic practice. All were implemented using rigorous item construction rules and processes to devise and refine items. This resulted in a seven-dimension GRS: Situation Awareness, History Gathering, Patient Assessment, Decision Making, Resource Utilization, Communication, and Procedural Skill.

Once the development phase was complete and a GRS was created, we subjected the scale to a quasi-experimental design. We recruited a range of clinicians (i.e., paramedic students at different levels of training and experienced paramedics) to complete the same case in a simulation setting, and then had raters, blinded to group, observe and rate the clinical

TABLE 4. Descriptive Statistics and Analysis of Variance Results by Dimension

	Descriptives					Analysis of Variance			
Dimension	Group	Mean	SD	95% CI		df	Mean Square	F	p-Value
Situation Awareness	1	2.36	1.24	1.85	2.87	2,78	52.49	32.44	0.00
	2	4.21	1.39	3.72	4.70				
	3	5.26	1.13	4.77	5.75				
History Gathering	1	3.10	1.28	2.57	3.63	2,78	10.94	9.03	0.00
	2	4.03	1.07	3.65	4.41				
	3	4.39	0.90	4.00	4.78				
Patient Assessment	1	2.16	1.02	1.74	2.58	2,78	36.00	25.09	0.00
	2	3.42	1.32	2.96	3.89				
	3	4.61	1.20	4.09	5.13				
Decision Making	1	2.28	1.30	1.74	2.82	2,78	57.11	30.00	0.00
	2	4.14	1.43	3.63	4.64				
	3	5.33	1.39	4.72	5.93				
Resource Utilization	1	2.78	1.00	2.37	3.19	2,78	26.06	22.34	0.00
	2	3.98	1.14	3.58	4.39				
	3	4.85	1.08	4.38	5.32				
Communication	1	3.42	1.19	2.93	3.91	2,78	8.06	5.14	0.01
	2	4.41	1.11	4.01	4.80				
	3	4.35	1.49	3.70	4.99				
Procedural Skill	1	2.78	1.44	2.18	3.38	2,78	35.46	21.40	0.00
	2	4.17	1.33	3.70	4.64				
	3	5.20	1.02	4.75	5.64				
OVERALL	1	2.30	1.22	1.80	2.80	2,78	46.27	29.48	0.00
	2	4.05	1.38	3.56	4.54				
	3	5.02	1.08	4.55	5.49				

CI = confidence interval; df = degrees of freedom; SD = standard deviation.

TABLE 5. Percentage of Individuals below Cut Score as Defined by the Modified Borderline Group Method (by Group and by Dimension)

	SA	HG	PA	DM	RU	COM	PS	Mean
Group 1	56%	56%	56%	64%	52%	48%	52%	54.9%
Group 2	12%	18%	21%	1%	12%	15%	15%	13.4%
Group 3	0%	1%	0.3%	0.3%	0.3%	18%	0.3%	2.9%

DM = Decision Making; G1 = group 1: novice-level paramedic students; G2 = group 2: entry-to-practice-level paramedic students; G3 = group 3: experienced active paramedics; COM = Communication; HG = History Gathering; PA = Patient Assessment; PS = Procedural Skill; RU = Resource Utilization; SA = Situation Awareness.

performances. Data collected from these processes allowed us to conduct item and reliability analyses and to evaluate the relationship of the scores to the training/experience level of the participants, a practice that has been successfully applied in other similar studies.[6,25,34,35,36]

The high interitem correlations observed suggest that the items, despite representing diverse dimensions, were possibly measuring a single construct.[36] These high interdimension correlations reinforce Lurie and colleagues' findings, which suggest that raters have difficulty differentiating between dimensions.[37] This may be an indication that, psychologically, raters form Gestalt categorical judgments about ratees as part of impression formation (i.e., a halo effect)[38] perhaps due to difficulty tracking multiple dimensions simultaneously.[33] Still, the scale demonstrated evidence of interrater and intrarater reliability, with minimal error attributed to rater. This may be in part due to deliberate efforts to align the label definitions (e.g., ready for independent practice) with the manner in which clinical supervisors naturally conceive of trainees' progress, a strategy that Crossley and colleagues have shown can improve rating practice.[39] Finally, the scores generated were significantly different between groups. These results strengthen confidence in the inferences made based on scores generated using this scale. That is, they strengthen the interpretive argument and suggest this GRS can be used for the assessment of paramedic clinical competence during entry-to-practice assessment processes.

These findings add to the broader health professional literature suggesting the use of GRSs to be a suitable measurement strategy. Crossley and Jolly claim that assessors judge performance more consistently and discriminatingly when not tied to process-level observation (i.e., reducing complex clinical performance to a series of individual steps).[40] This may help explain why the global ratings used in our study appear to be more reliable than those used by Han et al.,[14] despite the fact that we examined reliability within a group of novice practitioners just as they did. Our goal was to evaluate paramedic clinical competence at the entry-to-practice level. The variations in appropriate performance that can exist among clinicians at this level may not be amenable to process-level assessment (e.g., checklists). For instance, in making judgments regarding clinical performance, checklists may facilitate the assessment of occurrence (i.e., whether or not particular behaviors were present), but GRS may be more suitable for considering quality (i.e., how good the performance is) and suitability (i.e., whether or not the performance was good enough for entry to practice).[12] Rather, outcome- or structure-level assessments and definitions,[40] which are included in this GRS, may be better suited.

LIMITATIONS AND FUTURE RESEARCH

As always, there are limitations associated with this study. First, we used only one unscripted case (a medical cardiac patient). This tells us that performance can be reliably differentiated, but it limits external validity and prevents us from determining the extent to which individuals' general ability is captured by a single application of the scale. In terms of external validity, whether similar results would be found when assessing candidates attending to a trauma victim, for example, requires further study. That said, the case involved a number of interactions (e.g., communicating with unhelpful staff, integrating available resources, a need for selecting appropriate assessment strategies) that would likely be applicable in a variety of patient encounters. Further, the development phase of this study (i.e., task analysis, focus group, and expert panel) drew from a variety of contexts. Still, future studies will need to apply the GRS to other contexts (e.g., using different cases, in actual clinical competence examinations) to assess the generalizability of the results reported here. With respect to drawing inferences about individual paramedics' general level of competence, the universality of context specificity[40] requires that research be done to determine how many times the GRS needs to be applied (i.e., how many cases need to be observed) to generate stable representation of an individual's competence level.

Second, we had two independent expert raters score all 81 videos. The level of expertise of the raters as well as the repetition may have contributed to the results (e.g., high interrater reliability, low error variance), though it is worth noting that raters did not compare notes over the course of completing their assignment and, hence, were as prone to drifting apart in their perceptions as they were to come to a

mutual understanding of how well individual participants performed. Third, in critically appraising the GRS, we selected three groups (year 1 paramedic students, year 2 entry-to-practice students, and experienced paramedics) to test the scales' discriminative validity. The heterogeneity of these groups could be argued to have had an effect on our study results that demonstrate an ability to differentiate between groups. The scale was designed to support decisions for entry to practice and the groups selected, therefore, represent a range around the population of interest. The groups selected provide a range, therefore, around the population of interest. Further, within each group we were able to identify a range of performance levels (based on the high level of participant variance within group) and reasonable reliabilities were observed in all groups, including group 2, our intended target. This study employed a modified borderline group method and assumed a summative entry to practice setting. What cuts scores should be used in actual practice will depend on many factors, including the way in which the data are to be used (e.g., for formative or summative purposes) and the stakes involved in the decision to be made (e.g., academic progression or independent practice). Interested readers are directed to other sources for a more comprehensive treatment of assessment strategies and standard setting in health professions education.[2,12,41–43]

CONCLUSION

Paramedic program educators, employers, and certifying and/or licensing bodies all have a responsibility to ensure those who are ultimately given access to independent paramedic practice are indeed competent. This requires the use of appropriate process and measurement tools with sufficient scientific evidence to support inferences or interpretations based on the scores generated. This study provides support for use of our rigorously developed GRS in practice by demonstrating evidence of content validity, sound psychometric properties, limited construct irrelevant variance, and an ability to differentiate between levels of performance. Applied in the proper context, this scale could help strengthen decisions regarding paramedic clinical competence. Additional research is recommended to further support this interpretive argument, especially in other contexts.

References

1. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 suppl):S63–7.
2. Brennan RL. Educational Measurement. Westport, CT: Praeger Pub Text, 2006.
3. Regener H. A proposal for student assessment in paramedic education. Med Teach. 2005;27:234–41.
4. Lammers RL, Byrwa MJ, Fales WD, Hale RA. Simulation-based assessment of paramedic pediatric resuscitation skills. Prehosp Emerg Care. 2009;13:345–56.
5. Studnek JR, Fernandez AR, Shimber 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;18:1177–85.
6. Hodges B, Regehr G, McNaughton N, Toberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74:1129.
7. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73:993–7.
8. von Wyl T, Zuercher M, Amsler F, Walter B, Ummenhofer W. Technical and non-technical skills can be reliably assessed during paramedic simulation training. Acta Anaesthesiol Scand. 2009;53:121–7.
9. Martin M, Hubble MW, Hollis M, Richards ME. Interevaluator reliability of a mock paramedic practical examination. Prehosp Emerg Care. 2012;16:277–83.
10. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37:1012–6.
11. Martin J, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchinson C. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273–8.
12. Swanwick T. Understanding Medical Education. Wiley Online Library, 2010.
13. Govaerts MJB, van der Vleuten CPM, Schuwirth LWT. Optimising the reproducibility of a performance-based assessment test in midwifery education. Adv Health Sci Educ. 2002;7:133–45.
14. Han J, Kreiter CD, Park H, Ferguson KJ. An experimental comparison of rater performance on an SP-based clinical skills exam. Teach Learn Med. 2006;18:304–9.
15. Kane MT. Validation. In: Brennan RL (ed). Educational Measurement. 4th ed. Westport, CT: Praeger Publishers, 2006, pp 17–64.
16. Morgan P, Lam-McCulloch J, Herold-McIlroy J, Tarshis J. Simulation performance checklist generation using the Delphi technique. Can J Anesth. 2007;54:992–7.
17. de Villiers M, de Villiers P, Kent A. The Delphi technique in health sciences education research. Med Teach. 2005;27:639–43.
18. Graber M. Educational strategies to reduce diagnostic error: can you teach this stuff? Adv Health Sci Educ. 2009;14:63–9.
19. Ericsson K, Krampe R, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev. 1993;100:363–406.
20. Paramedic Association of Canada, National Occupational Competency Profile. Ottawa, Ontario, Canada: Paramedic Association of Canada, 2001.
21. Ontario Ministry of Health. Emergency Health Services Branch. BLS Standards of Care. Vol. 2.0. Toronto, Ontario, Canada: Publications Ontario, 2007.
22. Ontario Ministry of Health, Emergency Health Services Branch. Advanced Life Support Standards. Toronto, Ontario, Canada: Publications Ontario, 2007.
23. Ontario Ministry of Training Colleges and Universities, Paramedic Program Standards. Toronto, Ontario, Canada: Queens Printer for Ontario, 2008.
24. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1977.
25. Kim J, Neilpovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. Crit Care Med. 2006;34:2167.

26. Streiner D, Norman G. Health Measurement Scales: A Practical Guide to Their Development and Use. Fouth ed. New York: Oxford University Press, 2008.
27. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119:166, e7–16.
28. Brennan R Generalizability theory. Educ Measure Issues Pract. 1992;11(4):27–34.
29. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Med Educ. 2001;35:1043–9.
30. Humphrey-Murto S, MacFadyen JC. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. Acad Med. 2002;77:729–32.
31. Cizek GJ. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates Inc., 2001.
32. Kirwan BE, Ainsworth LK. A Guide to Task Analysis. Philadelphia: Taylor & Francis, 1992.
33. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ. 2012;Apr 7 (Epub ahead of print; doi: 10.1007/s10459-012-9370-3).
34. Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the MiniClinical Evaluation Exercise (MiniCEX). Acad Med. 2003;78:826–30.
35. Goff BA, Nielsen PE, Lentz GM, et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. Am J Obstet Gynecol. 2002;186:613–7.
36. Winckel CP, Reznick R, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. Am J Surg. 1994;167:423–7.
37. Lurie S, Mooney C, Lyness J. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. Acad Med. 2009;84:301–9.
38. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011;86(10 suppl):S1–7.
39. 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;45:560–9.
40. 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.
41. Cizek GJ. Bunch MB. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. Thousand Oaks, CA: Sage Publications, 2007.
42. Downing S, Yudkowsky R. Assessment in Health Professions Education. New York: Taylor & Francis, 2009.
43. Eva K. Assessment strategies in medical education. In: Salerno-Kennedy R, O'Flynn S (eds). Medical Education: State of the Art. Halifax, Nova Scotia, Canada: Nova Scotia Publishers, Inc., 2010.

APPENDIX 1. CASE SUMMARY

OVERVIEW

This case involved a paramedic (i.e., the candidate) working alone and responding to the side of a roadway for a patient with a decreased level of consciousness who was in the rear of a transfer company ambulance.* According to transfer company staff, the patients condition began with severe shortness of breath secondary to congestive heart failure that progressed into lethal arrhythmia and eventually cardiac arrest. The transfer company staff are "on scene" (i.e., in the rear of the transfer company vehicle with the patient) arguing over who is responsible for the current predicament.

CALL INFORMATION

Call for a 75-year-old male/female with shortness of breath.

CASE DETAILS

The patient [manikin] presented initially as responding only to painful stimuli with moans, was diaphoretic and tachypneic. His presenting rhythm was ventricular tachycardia. Presenting vital signs were: blood pressure: 68/48 mmHg, heart rate: 190 beats/min (ventricular tachycardia), respiratory rate: 30 breaths/min shallow and regular (crackles throughout), and blood sugar of 8.8 mmol/L. The patient had a history of Alzheimer's disease, coronary artery disease, two previous myocardial infarctions, congestive heart failure, cerebrovascular accident (no lasting deficits), hypertension, diabetes, and high cholesterol level. The medication list included Aricept, metoprolol, digoxin, lisinopril, Glucophage, and atorvastatin, and the patient was allergic to morphine.

*In Canada, some unregulated private companies may provide transfer services to patients. Generally, these unregulated transfer companies may not be held to the same standard as fully regulated ambulance services, including staff qualifications. Further, unregulated transfer companies are not authorized to transport patients directly to emergency departments.

APPENDIX 2. GLOBAL RATING SCALE FOR THE ASSESSMENT OF PARAMEDIC CLINICAL COMPETENCE

PARAMEDIC GLOBAL RATING SCALE

Candidate: _____ Rater: _____ Date: _____

Case Description: _____

Rating Label	Definition
1=Unsafe ……	Not performed as required. Performance compromised patient care / safety; serious remediation is required, unsuitable for supervised practice or progression.
2=Unsatisfactory …...	Performance indicated cause for concern. A potential for compromised patient care / safety exists; considerable improvement is needed. Not ready for supervised practice or progression.
3=Poor / Weak ……..	Inconsistently performed, and/or performance does not meet the standard, improvement is needed. More training / practice is needed before consideration for supervised practice or progression.
4=Marginal ……..…..	Occasionally performance is to standard, and/or performance meets minimum standards, improvement is recommended; suitable for supervised practice or progression with some remediation.
5=Competent …….	Often performed to standard, and/or performance is safe and to standard. Some areas could be improved. Ready for independent practice or progression with only minor concerns if any.
6=Highly Competent ..	Consistently performs to standard, and/or performance is safe and to standard. Occasionally exceeds the standard. Little improvement needed if any; ready for independent practice or progression.
7=Exceptional ……	Consistently demonstrates a high standard of performance, and/or consistently exceeds the standard enhancing patient safety; could be used as a positive example for others; highly recommended for independent practice or progression.

Situation Awareness	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to consider and integrate environmental, scene, resources and patient condition cues into the overall interaction, management and safety plan. This includes observing whole environment (all available data sources), anticipating likely events, discriminating between relevant and irrelevant data and avoiding tunnel vision (inappropriately focusing on elements to the exclusion of others). The individual is expected to demonstrate examples of situation awareness throughout the interaction and updating actions as necessary.

History Gathering	1	2	3	4	5	6	7
	UNSAFE	**UNSAT**	**POOR/WEAK**	**MARGINAL**	**COMPETENT**	**HIGHLY COMPETENT**	**EXCEPTIONAL**

Refers to the individual's overall ability to effectively and thoroughly gather an appropriate history (includes history of present illness and medical history) which is organized, appropriately structured, timed and focused according the clinical situation and level of urgency (context). This Includes interpreting and evaluating findings while discriminating between relevant and irrelevant findings. Also, refers to a demonstrated ability to include a consideration for differential diagnosis, while working toward a working diagnosis.

Patient Assessment	1	2	3	4	5	6	7
	UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL

Refers to the individual's overall ability to select and perform a physical exam and investigation of signs and/or symptoms that is organized and appropriate given the clinical situation and level of urgency. This includes interpreting and evaluating findings while discriminating between relevant and irrelevant findings. Also, refers to a demonstrated ability to continue appropriate reassessment / detailed assessment as needed. Finally, this also includes a consideration for differential diagnosis, while working toward a working diagnosis.

Decision Making	1	2	3	4	5	6	7
	UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL

Refers to the individuals overall ability to select an appropriate, safe, and effective management plan and/or strategy. Decisions should be based on and supported by findings, consideration of risks, benefits and differential diagnosis. This involves having adequate information for decisions made (i.e., avoiding premature closure) and ensuring decisions are appropriately prioritized, and timed. This also includes selecting an appropriate management device, method, or technique based on evidence (i.e., situation awareness, patient condition, resources etc) and context.

Resource Utilization	1	2	3	4	5	6	7
	UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL

Refers to the individual's overall ability to identify and use resources effectively to accomplish goals and maximize care. This includes the delegation of tasks, the coordination of efforts, selecting appropriate members (e.g., allied agencies, patients etc) for a given task, ensuring effectiveness and requesting additional resources as necessary. This also includes ability to function as a team with appropriate leadership.

Communication	1	2	3	4	5	6	7
	UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL

Refers to the individuals overall ability to clearly and accurately exchange information with the team, patient and/or bystander for optimal patient care and team effectiveness. This includes the use of concise and appropriate language, ensuring statements are directed at appropriate individuals and that messages are heard / received (i.e., closes the loop). This also includes demonstrating effective listening skills, demonstrating empathy, responding appropriately to statements by the team, patient or bystander. Actions are appropriately communicated with team, patient and bystander. Verbal and non-verbal are appropriate and congruent.

Procedural Skill	1	2	3	4	5	6	7
	UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL

Refers to the individuals overall ability to complete psychomotor or procedural skills or tasks effectively, appropriately and to standard. This involves a familiarity with equipment used, ensuring appropriate and safe application while completing tasks to standard and avoiding commission or omission errors. This also involves adaptability to failures / problems (as necessary) and ensuring team, patient and bystander safety while performing these procedures; includes appropriate execution, properly sequenced, and evaluating / reevaluating effectiveness.

Overall Clinical Performance							
UNSATISFACTORY			SATISFACTORY				
1	2	3	4	5	6	7	
UNSAFE	UNSAT	POOR/WEAK	MARGINAL	COMPETENT	HIGHLY COMPETENT	EXCEPTIONAL	