Received: 13 June 2024 | Revised: 19 December 2024 | Accepted: 22 December 2024

DOI: 10.1002/aet2.11061

ORIGINAL ARTICLE

Validity evidence of a resuscitation team leadership assessment measure for use in actual trauma resuscitations

Elizabeth D. Rosenman MD[1] | James A. Grand PhD[2] | Rosemarie Fernandez MD[3]

[1]Department of Emergency Medicine, University of Michigan Medical School, Seattle, Washington, USA

[2]Department of Psychology, University of Maryland, College Park, Maryland, USA

[3]Department of Emergency Medicine, University of Florida, Gainesville, Florida, USA

Correspondence
Elizabeth D. Rosenman, Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA.
Email: roseeliz@med.umich.edu

Funding information
Agency for Healthcare Research and Quality, Grant/Award Number: 1R18HS022458-01A1; U.S. Department of Defense, Grant/Award Number: W81XWH-18-1-0089

Abstract

Background: Team leadership is a critical skill in trauma resuscitation teams, linked to better teamwork and improved patient care. There are numerous published team leadership assessments, though data regarding the performance of these measures in patient care settings (vs. simulation-based settings) remain limited. There remains a need for a valid, reliable, and efficient measure of resuscitation team leadership in the clinical setting to support medical education and research efforts.

Methods: We constructed a 12-item behaviorally anchored rating scale (BARS) to measure trauma team leadership. Multiple raters then used the BARS to measure team leadership in 360 recorded trauma resuscitations across 60 participants. In addition to examining inter-rater reliability, we examined the construct validity of the BARS assessment through both correlational and latent modeling techniques to compare the ratings collected with the BARS to those collected using a previously studied checklist-based assessment using a multitrait–multimethod (MTMM) approach. Lastly, we examined the criterion validity of the BARS measure by examining its relationship with previously obtained patient care scores.

Results: BARS items demonstrated high inter-rater reliability when scores were computed using observations averaged over multiple raters (mean item intraclass correlations ICC1k0.90, item range0.85–0.98). The correlation between the aggregate ratings from the team leadership BARS and checklist measure demonstrated a strong positive correlation ($r=0.75$), and the MTMM analyses indicated consistent evidence for both convergent (mean monotrait–heteromethod $r=0.50$) and discriminant (mean heterotrait–heteromethod $r=0.27$) validity. Hierarchical Bayesian regression analyses revealed that aggregate BARS scores were predictive of patient care scores ($\beta=7.06$, 95% HDI 3.76–10.43).

Conclusions: The team leadership BARS and a previously studied checklist-based team leadership measure produced convergent assessments of team leadership behavior in the present data. Furthermore, higher overall ratings on the BARS correlated with better patient care delivery at the team level.

KEYWORDS
assessment, measure validation, resuscitation, team leadership

Supervising Editor: Sally Santen

INTRODUCTION

Team leadership is an important skill in health care resuscitation teams, including emergency trauma teams. Effective team leadership is linked to better teamwork[1] and improved patient care delivery.[2] The important role of team leadership makes it a potential target for patient safety initiatives; however, a robust approach to measuring team leadership is necessary to support training and research aimed at improving resuscitation team leadership.

Resuscitation team leadership measurement remains challenging despite numerous published assessment measures.[3] This is due, in part, to the complex resuscitation teams and clinical environments in which team leaders function and the interdependency of the team leader with the team and environment.[4] Team and team leader interdependence is especially challenging, as it has proven difficult to assess both within, and beyond, health care.[5–7] It is therefore not surprising that many team leadership measures have been developed and/or tested using simulation-based care as stimuli, which provides a degree of standardization and control.[3,8]

Published resuscitation team leadership assessments are mixed in approach, with global rating scales (GRSs) and checklists most commonly used.[3] Existing research evaluating the psychometric properties of these two assessment approaches provides conflicting recommendations. Inter-reliability is cited as being high in both checklist measures and GRSs in simulated settings.[9] Very little is known, however, about the relative performance of these assessment approaches in the clinical setting, which introduces more complexity and more variability. Checklists are often tailored, by necessity, to a very specific event or context. Developing a new checklist for each setting can be time-consuming and makes it difficult to amass a body of validity evidence over time and across contexts. Conversely, GRSs are often easier to implement but carry their own limitations. Most notably, GRSs may be prone to rater biases that can limit reliability and validity.[10,11] Relying on the expertise of the evaluator may restrict the available pool of evaluators and/or require extensive rater training.

A potential solution to the challenges with these two approaches, is the use of behaviorally anchored rating scales (BARS).[12,13] BARS utilize specific, observable behaviors that can promote reliability and facilitate feedback, while using a smaller number of items that retains a more holistic approach to the assessment. However, prior work has demonstrated that assessment measures, including a team leadership BARS, found to be highly reliable when applied to simulation-based care, do not always perform as well when applied to actual patient care.[14,15]

To address the persistent gap in trauma team leadership assessment in the clinical setting, we developed a trauma team leadership assessment measuring using a BARS approach. The measure was specifically intended for application to actual trauma resuscitations and evaluation of the measure includes comparing it to a patient care activity measure and to another, previously developed, checklist-based team leadership assessment with supporting evidence of validity.

METHODS

Study overview

We developed a trauma team leadership BARS based on an existing detailed trauma team leadership behavioral coding scheme with supporting validity evidence.[2] The BARS was refined through subject matter expert (SME) input and evaluated using a collection of recorded trauma resuscitations. These same resuscitations were previously scored for clinical performance and for team leadership using checklist-based assessments. Evidence of validity is presented using Messick's framework,[16] as described by Downing.[17] The University of Washington Institutional Review Board approved this study.

Measure development

We developed a trauma team leadership BARS using a previously published checklist-based assessment as the foundation.[2] This checklist was chosen because it was designed specifically for trauma resuscitations and demonstrated supporting validity evidence when applied to actual patient care. Target leadership behaviors included in the checklist were identified using a taxonomy of leadership behaviors,[18,19] two systematic reviews of resuscitation team leadership,[3,20] and a trauma team leadership training curriculum.[21] To translate this work into a BARS assessment, the research team, with expertise in team science and emergency department trauma care, used the checklist to systematically identify a list of observable leadership behaviors expected to occur during trauma resuscitations. These behaviors were used to create descriptive and diagnostic rating anchors. Collectively, this foundational work supports the content validity of the BARS.[17]

In several instances, checklist items were aggregated and/or used to inform the anchors used for a BARS item's rating scale. Additionally, two new items were introduced to the BARS to improve its generalizability. First, a new item labeled *asserts control* was added. A similar behavior was reflected in the prior checklist measure to assess whether leaders explicitly claimed a leadership role near the beginning of patient care (*assumes team leadership role*). The additional BARS item was added to capture these behaviors throughout a resuscitation, rather than limiting it to the start of the resuscitation. Second, a new item labeled *huddle: overall quality* was added. The BARS and prior checklist included items intended to capture specific components of huddle behavior (i.e., *engagement* and *content*). Expert input suggested that there would be value in capturing an overall assessment on these high-impact communications events as well.

External SMEs provided feedback on the relevance, importance, and clarity of the individual items and anchors. SMEs included team science experts ($n = 2$), with over 25 years of combined experience in leadership science and health care assessment

design, and board-certified emergency physicians and medical educators ($n=2$), with over 10 years of combined experience in trauma care and medical education. Where expert opinion indicated that specific aspects of the team leadership measure were confusing, or were not sufficiently representative, items were revised and reevaluated. The measure was then piloted on video-recorded trauma resuscitations by these same SMEs as well as the research team. Where anchor behaviors could not be clearly observed, or were not clearly interpreted, items were revised. Expert input and pilot testing the items further supports the content validity of the measure. The final BARS assessment consisted of 12 items reflecting eight facets of leader behavior identified as relevant to trauma team leadership. The rating scale for each item uses a 3-point scale whose anchors provide descriptions of specific behaviors reflecting poor (1), average (2), or excellent (3) team leadership on the targeted dimension as determined by SMEs. Table 1 summarizes the final set of leadership dimensions targeted by the assessment and the relevant BARS and checklist items used to measure each construct. The complete BARS is available in the supplemental material (Table S1).

Primary data

The team leadership BARS was applied to a repository of 360 audiovisual recordings of actual trauma resuscitations. The recordings were gathered from 60 participants, with each participant observed across six different resuscitations. These recordings were acquired from a Level I trauma center at a large, urban, academic center with a tiered trauma response system based on patient acuity. Inclusion

TABLE 1 Leadership concepts targeted by the checklist and BARS items.

Concepts	Checklist Items	BARS Item
Assuming and maintaining leadership	Explicit assumption of leadership	Assumes leadership role
		Asserts control
Briefing team prior to patient arrival ("prebrief")	Explicitly engages team in a prebrief	Prebrief: Engagement
	Content of prebrief: • Info sharing • Planning • Role assignment • Seeks input	Prebrief: Content
Briefing team immediately after patient arrival to update/modify plan ("rebrief")	Highlights updates from prehospital report	Rebrief
	Communicates impact of new information on plan/role assignment or priorities	
Briefing team during ongoing patient care ("huddle")	Explicitly engages team in a huddle	Huddle: Engagement
	Content of huddle: • Info sharing • Planning • Role assignment • Seeks input	Huddle: Content Huddle: Overall quality
Sharing information with team	Summarizes patient status/results	Information sharing
	Provides interpretation of the facts for the team	
Creating or updating a plan	States plan with future steps	States plan
	States plan with priorities	
Assigning a task or role to team member	Identifies individuals for tasks	Role assignment
	Factors in team member skill set and/or requests a check back when task is complete	
Seeking team input on ideas or potential barriers	Asks for team input	Seeks input
	Asks for potential delays or barriers	

Abbreviation: BARS, behaviorally anchored rating scales.

and exclusion criteria for recording acquisition have been previously published.[3] The recordings included two views (foot of bed and side view). Recordings were coded using Noldus Observer XT software.

Eight critical care technicians were recruited to perform team leadership ratings using the BARS assessment. Rater training included reviewing a document that provided descriptions and examples for each item, followed by coding two resuscitations and comparing and discussing scores. Each video was then independently coded by two raters. The unreconciled scores were used for determining interrater reliability. Discrepancies in coding were then identified and reconciliation was performed for items in which the scores differed by more than one point (i.e., one rater coded "poor" and one rater coded "excellent"). This was done to improve the accuracy and reliability of the final scores used for analyses comparing the BARS to other performance measures. Reconciliation was performed by an emergency medicine physician with expertise in team leadership and trauma care.

Secondary data used for validity comparison

Secondary data, including leadership scores and patient care scores, were obtained using the same 360 audiovisual recordings described above and have been previously published.[3] For the comparison leadership scores, a checklist-based metric was used by a separate group of trained raters to code all team leader behaviors. A random sampling of 40% ($n = 144$) of the observations were coded in duplicate. Prevalence is a known problem in observational codes targeting low base rate events. Following recommendations by Byrt et al.,[22] we calculated the probability and bias adjusted kappa (PABAK), which was 0.97 across all items. The BARS scores were compared to individual item and composite scores from the checklist measure to demonstrate convergent and discriminant validity (relationship to other variables).[16,17] Convergent validity is a measure of how closely two instruments that are intended to measure the same construct are related to each other. Discriminant validity is a measure of whether two instruments, that are not intended to measure the same construct, are not closely correlated.

For the patient care scores, a previously published patient care measure was used and coding was performed by a separate group of trained raters.[2,23] The measure was developed using the Advanced Trauma Life Support curriculum[24] and previously published trauma care checklists in simulated[25] and live[26–29] patient care. The purpose of the measure is to quantify the quality of early trauma resuscitation efforts, taking into consideration the clinical variability that can impact the relevance of certain patient care tasks (e.g., blood transfusions, tube thoracostomies). A random sampling of 10% ($n = 36$) of the observations were coded in duplicate, with an average Cohen's $\kappa = 0.8$ (SD ± 0.09). The BARS scores were compared to composite patient care scores for the purpose of demonstrating test–criterion correlations (relationship to other variables).[16,17]

Data analysis

Data preparation

Multiple checklist items were identified as relevant to assessing five out of the eight leadership dimensions measured on the BARS (see Table 1). To facilitate comparisons between the BARS and checklist measure, and when examining the multitrait–multimethod (MTMM) analyses, item parcels were created by averaging together those checklist items to create a composite score. For example, two checklist items were used to capture *information sharing*, including "summarizes patient status/results" (e.g., "The blood pressure is 75/40" and "provides interpretation of the facts for the team") (e.g., "Because the patient is hypotensive I am worried about hemorrhagic shock"). These items were combined to create a single "information sharing" parcel for this dimension on the checklist measure.

Inter-rater reliability

Inter-rater reliability was evaluated by computing intraclass correlation coefficients (ICCs) for each item on the BARS prior to reconciliation.[30] Because the same pair of coders did not rate all observations, the ICC1 and ICC1k coefficients are most appropriate for these analyses. ICC1 provides information about the reliability of a rating based on the assumption that the final scores from a measure are computed using the responses of a single rater. In contrast, ICC1k provides information about the reliability of a rating based on the assumption the final scores from a measure are computed by averaging the responses of multiple raters. The ICC calculations were conducted in R using the *psych* package.[31,32]

MTMM analyses

MTMM designs entail collecting measures of multiple traits (e.g., leadership behavior) using multiple methods (e.g., checklist vs. BARS). MTMM approaches are commonly used as a means for examining the construct validity of a measurement tool by evaluating the extent to which conclusions about the traits expressed/possessed by a focal unit (i.e., a leader) are corroborated across multiple assessment techniques.[33] Two analytical strategies were followed to perform the MTMM analyses (Table 2). Unlike the reliability analyses, both MTMM analyses relied on the reconciled BARS data to minimize the potential of rater error when interpreting the validation results. First, simple zero-order correlations were computed among the leadership items from the BARS and those from the behavioral checklist to examine evidence of convergent validity (monotrait–heteromethod correlations), discriminant validity (heterotrait–heteromethod correlations), and methods effects (heterotrait–monomethod correlations).

TABLE 2 Overview of the MTMM analyses performed in this study comparing a team leadership assessment BARS to a preexisting team leadership assessment checklist.

MTMM analyses		Explanation	Purpose and summary of results
Correlations	Monotrait–heteromethod	Correlations from measures of the same construct attained using different measurement methods (e.g., correlation between measures of huddle engagement in the BARS and checklist data)	Strong correlations provide evidence of convergent validity Mean $r=0.50$
	Heterotrait–heteromethod	Correlations from measures of different constructs attained using different measurement methods (e.g., correlation between measures of huddle engagement and role assignment in the BARS and checklist data, respectively)	Weak correlations provide evidence of discriminant validity Mean $r=0.27$
	Heterotrait–monomethod	Correlations from measures of different constructs attained using the same measurement method (e.g., correlation between measures of huddle engagement and role assignment in the BARS data)	Strong correlations provide evidence of method effects (i.e., extent to which measurement method is likely to contribute to systematic measurement error) Mean BARS $r=0.43$ Mean checklist $r=0.26$
CFA	Model 1[a] vs. Model 2	Compare baseline model to alternative model without the latent leadership trait factors so only latent method factors remains. A significant chi-square difference test indicates the baseline model (Model 1) fits the data better than the alternative model.	Comparison of model fit provides evidence of whether measure provides meaningful information about constructs of interest (i.e., convergent validity) $\Delta\chi^2(51)=1933.9$, $p<0.001$
	Model 1[a] vs. Model 3	Compare baseline model to alternative model with latent leadership traits perfectly correlated while retaining distinct latent method factors. A significant chi-square difference test indicates the baseline model (Model 1) fits the data better than the alternative model.	Comparison of model fit provides evidence of whether constructs of interest can be meaningfully differentiated from one another (i.e., discriminant validity) $\Delta\chi^2(28)=871.3$, $p<0.001$
	Model 1[a] vs. Model 4	Compare baseline model to alternative model with latent method factors perfectly correlated while retaining distinct latent leadership trait factors. A significant chi-square difference test indicates the baseline model (Model 1) fits the data better than the alternative model.	Comparison of model fit provides evidence of whether measurement method affects measurement of the constructs of interest (i.e., methods effect) $\Delta\chi^2(2)=293.5$, $p<0.001$

Abbreviations: BARS, behaviorally anchored rating scale; CFA, confirmatory factor analysis; MTTM, multitrait–multimethod.

[a]The baseline model for the CFA (Model 1) is available in the supplemental material (Figure S1).

Second, convergent and discriminant validity of the BARS measure were also examined using a confirmatory factor analysis (CFA) MTMM approach to account for the latent measurement models of both assessments and their potential method effects.[34] Data from the BARS and checklist were first fit to a baseline model that modeled items loading onto distinct latent leadership traits and measurement methods (Model 1). Model fit between the data and this baseline model were examined using standard indices and criteria (e.g., CFI >0.90, RMSEA <0.08, SRMR <0.05; patterns of significance for factor loadings).[35] A series of alternative models were then constructed and statistically compared to the baseline model using chi-square difference tests. The first alternative model (Model 2) removed the latent leadership trait factors and retained only the latent methods factors to examine whether the data collected with the assessments reflect distinct traits or are completely captured by measurement method (convergent validity). The next alternative model (Model 3) permitted the latent leadership traits to be perfectly correlated while retaining distinct latent methods factors to examine whether the assessments captured a single/general latent trait (i.e., leadership) versus multiple/distinct latent traits (discriminant validity). The final alternative model (Model 4) permitted the underlying latent methods factors to be perfectly correlated while retaining the distinct latent leadership trait factors to examine whether the different measurement methods capture unique information about the observed items (methods effect). The CFAs were conducted in R using the *lavaan* package.[36]

Criterion validity

Lastly, the predictive validity of the team leadership BARS was assessed by examining its relationship with patient care scores (correlation with other variables).[17] Given the nested nature of the data (i.e., multiple observations of leadership and patient care per participant), parameter estimates were computed using a two-level random effects Bayesian regression model to account for nonindependence between observations. Given the absence of compelling previous evidence to accurately inform the choice of priors for our model, we used weakly informative priors for all model parameters so as not to "overwhelm" the observed data or unduly influence final model estimates.[37] The primary relationship of interest is the pooled linear relationship between scores on the team leadership BARS and patient care for leaders. We thus evaluated and reported the median and 95% highest-density interval from the posterior distribution for this parameter. Analyses were conducted in R using the *rstanarm* package with default settings for setting model priors.[38] Where reported, p-values ≤0.05 were used to interpret statistical significance.

RESULTS

Composite scores on the team leadership BARS (i.e., a leader's average rating across all BARS items) ranged from 1–3 with a mean of 1.42 (SD ± 0.41). Variables for comparison, including the checklist-based scores for team leadership and patient care, existed for all 360 resuscitations.

Inter-rater reliability

Table 3 reports the ICC1 and ICC1k values for each BARS item prior to reconciliation. The observed ICC1k values reflect excellent inter-rater agreement (ICC > 0.75)[39] for all BARS items (mean 0.90, range across dimensions 0.85–0.98). In contrast, the observed ICC1 values demonstrated fair agreement (ICC 0.40–0.59)[39] for most items (mean 0.56, range 0.41–0.87).

TABLE 3 Inter-rater reliability results for team leadership BARS items.

BARS Item	ICC1	ICC1k
Assumes leadership role	0.57	0.91
Asserts control	0.42	0.85
Prebrief: Engagement	0.87	0.98
Prebrief: Content	0.81	0.97
Rebrief	0.43	0.86
Huddle: Engagement	0.59	0.92
Huddle: Content	0.51	0.89
Huddle: Overall quality	0.59	0.92
Information sharing	0.53	0.90
States plan	0.57	0.91
Role assignment	0.41	0.85
Seeks input	0.42	0.85

MTMM correlation analysis

The overall correlation between aggregate scores on the team leadership BARS and the checklist-based team leadership measure (i.e., sum of all checklist items) was $r = 0.75$ ($p < 0.001$), indicating a high degree of correspondence between the two measures. An examination of the correspondence between these measures at the item level is presented in Table 4, which shows the MTMM matrix summarizing the zero-order correlations among the observed ratings collected using the two different measures.

Overall, the MTMM correlations indicate evidence of both convergent (mean monotrait–heteromethod correlation of 0.50) and discriminant (mean heterotrait–heteromethod correlation of 0.27) validity (Table 2). However, there were six dimensions measured on the BARS (*asserts control*, *huddle: engagement*, *huddle: content*, *information sharing*, *role assignment*, and *seeks input*) for which its observed monotrait–heteromethod correlation was exceeded by at least one of its corresponding heterotrait–heteromethod correlations, indicating comparatively weaker convergent validity evidence for these items. In some such instances, these BARS scores exhibited a stronger heterotrait–heteromethod correlation with a checklist item that captured a highly related concept (e.g., the BARS items for *huddle: engagement* and *huddle: content* shared stronger heterotrait–heteromethod correlations with the *huddle: overall quality* checklist item). In other instances, these BARS items exhibited a stronger heterotrait–heteromethod correlation with a checklist item that may suggest raters relied on similar sources of information for rating the team leader (e.g., the BARS item for *information sharing* shared its strongest heterotrait–heteromethod correlation with the *huddle: overall quality* checklist item, potentially indicating that raters using the BARS may have been partially relying on their perception of a team leader's huddle when assessing that leader's information sharing). Nevertheless, the overall pattern of correlations generally supports the convergent and discriminant validity of raters' assessments of team leadership behaviors using the BARS with those measured using the checklist.

MTMM CFA

The factor loadings from the baseline CFA MTMM model (Model 1) are available in the supplemental material (Figure S1), and Table 5 summarizes the model fit statistics and model comparison results for all four CFA models computed. All but one of the trait-to-item factor loadings achieved statistical significance, and evidence for the overall fit of the MTMM data to the baseline model received moderate support based on conventional standards

TABLE 4 MTMM correlation matrix between the BARS and the behavioral checklist ($n = 360$).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Assumes leader role (B)	—																						
2. Asserts control (B)	0.56	—																					
3. Prebrief: Engagement (B)	0.49	0.34	—																				
4. Prebrief: Content (B)	0.42	0.42	0.75	—																			
5. Rebrief (B)	0.32	0.39	0.25	0.26	—																		
6. Huddle: Engagement (B)	0.30	0.52	0.17	0.24	0.23	—																	
7. Huddle: Content (B)	0.27	0.53	0.22	0.24	0.27	0.68	—																
8. Huddle: Overall quality (B)	0.30	0.58	0.23	0.27	0.31	0.75	0.87	—															
9. Information sharing (B)	0.47	0.65	0.43	0.48	0.38	0.65	0.67	0.74	—														
10. States plan (B)	0.44	0.65	0.37	0.43	0.32	0.60	0.64	0.66	0.71	—													
11.Role assignment (B)	0.30	0.53	0.33	0.45	0.21	0.37	0.52	0.48	0.52	0.55	—												
12. Seeks Input (B)	0.25	0.38	0.40	0.50	0.29	0.25	0.35	0.37	0.37	0.35	0.42	—											
13. Assumes leader role (C)	0.52*	0.31*	0.27	0.20	0.26	0.22	0.19	0.19	0.29	0.25	0.13	0.11	—										
14.Prebrief: Engagement (C)	0.30	0.25	0.66*	0.56	0.21	0.19	0.24	0.24	0.38	0.33	0.26	0.38	0.22	—									
15. Prebrief: Content (C)	0.29	0.30	0.61	0.69*	0.20	0.24	0.24	0.27	0.45	0.40	0.33	0.35	0.24	0.80	—								
16. Rebrief (C)	0.26	0.24	0.14	0.07	0.58*	0.19	0.12	0.17	0.26	0.20	0.06	0.05	0.24	0.12	0.11	—							
17.Huddle: Engagement (C)	0.33	0.45	0.26	0.27	0.15	0.73*	0.51	0.60	0.56	0.52	0.29	0.18	0.28	0.25	0.26	0.15	—						
18. Huddle: Content (C)	0.10	0.20	0.16	0.15	0.20	0.25	0.29*	0.28	0.23	0.27	0.26	0.24	0.11	0.21	0.25	0.08	0.25	—					
19.Huddle: Overall quality (C)	0.32	0.45	0.29	0.29	0.19	0.73	0.56	0.63*	0.58	0.55	0.34	0.24	0.28	0.29	0.29	0.15	0.96	0.46	—				
20. Information sharing (C)	0.16	0.28	0.15	0.19	0.16	0.35	0.33	0.33	0.39*	0.39	0.21	0.16	0.20	0.21	0.24	0.13	0.33	0.37	0.39	—			
21. States plan (C)	0.27	0.44	0.19	0.24	0.24	0.35	0.41	0.41	0.41	0.62*	0.37	0.23	0.24	0.22	0.27	0.20	0.36	0.27	0.40	0.41	—		
22. Role assignment (C)	0.12	0.16	0.20	0.20	0.12	0.22	0.22	0.20	0.21	0.26	0.25*	0.11	0.10	0.21	0.23	0.05	0.25	0.44	0.31	0.15	0.31	—	
23. Seeks Input (C)	0.17	0.25	0.19	0.22	0.20	0.23	0.26	0.26	0.27	0.28	0.22	0.32*	0.04	0.21	0.22	0.06	0.18	0.47	0.29	0.21	0.25	0.24	—

Note: (B) = item from behaviorally anchored rating scale (BARS). (C) = item from behavioral checklist. Values below the dashed vertical line correspond to the heterotrait–monomethod correlations for items measured using the behavioral checklist. $p < 0.05$ for all correlations greater than 0.10.

Abbreviations: BARS, behaviorally anchored rating scale; MTMM, multitrait–multimethod.

*Monotrait-heteromethod correlations.

TABLE 5 Model fit and comparison statistics of MTMM confirmatory factor analysis.

Model	χ^2	*df*	RMSEA	SRMR	CFI	AIC	Model comparison	$\Delta\chi^2$	Δdf
M1: Correlated traits, correlated methods (baseline MTMM model)	577.1*	178	0.08	0.07	0.93	10,797	—	—	—
M2: No traits, correlated methods	2511*	229	0.17	0.15	0.60	12,629	M1 vs. M2	1933.9*	51
M3: Single trait, correlated methods	1448.4*	206	0.13	0.10	0.78	11,612	M1 vs. M3	871.3*	28
M4: Correlated traits, single method	870.6*	180	0.10	0.07	0.88	11,087	M1 vs. M4	293.5*	2

Note: When comparing any two models, a better fitting model is generally indicated by a lower chi-square, lower RMSEA, lower SRMR, lower AIC, and higher CFI. Model comparison M1 versus M2 provides evidence of convergent validity, M1 versus M3 evidence of discriminant validity, and M1 versus M4 evidence of methods effects.

Abbreviations: AIC, Akaike information criterion; CFI, comparative fit index; MTMM, multitrait–multimethod; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

*$p < 0.001$.

($\chi^2[178] = 577.1$, $p < 0.001$; CFI 0.93, RMSEA 0.08, SRMR 0.07). Of greater importance to assessing the convergent and discriminant validity of the team leadership BARS is the comparison between the baseline model and the alternative models shown in Table 5. The baseline model fit the MTMM data significantly better than models in which variance on the BARS and checklist items is accounted for by only their corresponding measurement method (i.e., no latent traits modeled, Model 1 vs. Model 2) or a model in which the measures are considered unidimensional (i.e., all items load on a single latent trait, Model 1 vs. Model 3). Although a positive correlation between the latent BARS and checklist methods factors was observed in the baseline model ($r = 0.36$), a model in which separate methods effects were not included failed to improve model fit (i.e., no separate BARS or checklist methods factors, Model 1 vs. Model 4).

Overall, the results of the CFA MTMM analyses provide evidence that assessments of team leadership using the BARS and behavioral checklist capture multiple and comparable dimensions of leader behavior. Consistent with the correlational MTMM analyses, the findings also indicate that the BARS and checklist measures appear to exhibit similar but non-identical methods effects. In sum, these findings indicate that the BARS and behavioral checklist measures both produced convergent assessments of team leadership in the present data.

Criterion validity

The median posterior estimate for the relationship between leaders' composite score on the team leadership BARS and patient care was 7.06 (95% HDI 3.76–10.43); median estimate for the model intercept was 51.82 (95% HDI 46.83–56.7). These results indicate that teams whose leaders received higher overall leadership ratings on the BARS tended to have higher patient care scores at the team level.

DISCUSSION

This work provides evidence of the psychometric quality for a team leadership BARS. Evidence of validity presented here includes content validity, internal structure, and relationship to other variables.[17] The analyses we present indicate that the BARS measure can be used to generate reliable ratings of team leadership in real trauma resuscitations. Furthermore, the MTMM analyses indicate that assessments of team leadership using the BARS measure demonstrate very good convergent and discriminant validity when compared to a previously validated team leadership checklist measure applied to the same data set. Finally, teams with leaders who received higher overall ratings on the BARS tended to deliver demonstrably better patient care. Evaluating for, and demonstrating, this type of criterion validity can be very challenging due to the complexity of the clinical environment and the interdependence between team leaders and teams.[4,40] These findings are significant because they indicate that the BARS tool can be used to accurately assess multiple and distinguishable facets of trauma team leadership using nonexpert raters and a more holistic approach that does not require coding every verbalization related to team leadership.

With respect to the psychometric properties of the team leadership BARS, the observed ICC1k values indicated that strong interrater reliability was achieved for each of the BARS items among our sample of raters. These results indicate that using multiple raters and subsequently averaging their ratings together would be expected to result in highly reliable measurements. In contrast, the observed ICC1 values for most of the BARS items failed to reach standard levels of acceptability. These results suggest that relying on only a single rater's assessments using the team leadership BARS would be expected to result in a larger degree of measurement error than is typically desired for either research or training purposes. Although the nature of the computations for the ICC guarantee that ICC1k values will always be higher than those for ICC1,[30] this difference was quite large for some of the BARS items (e.g., *asserts control* and *assign*

roles). This argues against using the leadership BARS, as currently presented, with a single rater in high-stake evaluations. This could be addressed by using multiple raters, additional rater training, and/or modifying the most challenging items. It is also important to remember that the nonreconciled data were used for these analyses so as not to capitalize on rater consensus when interpreting the reliability of the BARS; consequently, the ICC results can be considered more conservative estimates and on the lower range of the reliability that might be expected when using the assessment tool.

The correlations presented in Table 4 also provide insight into potential methods effects in the team leadership BARS that should be considered when using and interpreting ratings. The observed heterotrait–monomethod correlations shown in the matrix suggest that assessments taken using the BARS may be prone to measurement errors that make it challenging for raters to distinguish between certain dimensions of leadership (e.g., halo errors). Such patterns are common with measurement tools that use rating scales[9,10] and could potentially be improved by expanding the number of rating points and anchors for each dimension. Relatedly, the higher magnitude of inter-item correlations with the team leadership BARS compared to the checklist is also not surprising given that the former asked raters to provide retrospective ratings on a relatively small number of dimensions at the end of an observation, whereas the latter asked raters to record their observations in real time across a large number of specific behaviors. The heterotrait–monomethod correlations for the team leadership BARS revealed that the set of items most prone to a potential method effect were the three items assessing "huddles" and those related to asserting control, sharing information, stating plans, and assigning roles (mean $r=0.61$). Notably however, this same group of items also exhibited the strongest heterotrait–monomethod correlations for the checklist data (mean $r=0.38$) and among the heterotrait–heteromethod correlations (mean $r=0.37$). Taken together, these patterns suggest that huddles are highly salient leadership events to raters—irrespective of measurement method—and that many of the rated leadership behaviors may happen (or not) during these times.

This work is a part of a larger body of research attempting to refine team leadership assessment in resuscitation teams using a functional model of leadership, as described by Morgeson et al.[41] In our prior work we have looked specifically at EM residency training, including comparing assessment reliability in simulation versus live patient care[14] and reviewing team leadership assessment practices relative to the Accreditation Council for Graduate Medical Education requirements.[18] The work that most directly informed this current study was focused on interdisciplinary trauma resuscitations in the clinical setting.[2] Ultimately, the goal of this work is to develop a behavioral measure of leadership that can facilitate real-time assessments during clinical care (e.g., by peer or attending physicians on shift), simulation-based assessments (e.g., medical educators during direct or recorded observation), and team leadership–focused research efforts (e.g., larger sample sizes with a potentially more diverse group of raters). The results presented here advance our prior work by presenting validity evidence for using this team leadership BARS to assess a relatively large number of actual clinical events using nonexpert raters. Further work is needed to evaluate the validity evidence of this measure when applied in real time at the bedside.

LIMITATIONS

There are several limitations to this study. First, rater observations were conducted by video review. This allowed for unobtrusive direct observation; however, it also gave raters the opportunity to pause and replay the recordings. The performance of this assessment measure, when applied in real time at the bedside is unknown. Second, this work focused on emergency trauma resuscitations from a single, academic institution. Finally, the raters used for the BARS had basic medical knowledge but were not content experts in team leadership or emergency trauma care. Although they did undergo rater training, it is unclear how this measure would have performed if a more experienced pool of raters was used.

CONCLUSIONS

Team leadership will always be challenging to assess in the clinical environment; however, this work supports using a BARS to assess team leadership reliably and accurately while being mindful of resources. Whether for research or medical education, rigorous assessment requires identifying, training, and maintaining an appropriate pool of raters, which can be time-consuming and expensive. The trauma team leadership BARS presented here was easily implemented by novice raters with basic rater training. Future work is necessary to evaluate the use of this trauma team leadership BARS at the bedside.

AUTHOR CONTRIBUTIONS

Study concept and design: Elizabeth D. Rosenman, Rosemarie Fernandez. Acquisition of data: Elizabeth D. Rosenman, Rosemarie Fernandez. Analysis and interpretation of the data: Elizabeth D. Rosenman, James A. Grand, Rosemarie Fernandez. Drafting of the manuscript: Elizabeth D. Rosenman. Critical revision of the manuscript: James A. Grand, Rosemarie Fernandez. Statistical expertise: James A. Grand. Acquisition of funding: Rosemarie Fernandez.

FUNDING INFORMATION

Funding and support for this project was provided by the Agency for Healthcare Research and Quality (1R18HS022458-01A1) and the Department of Defense (W81XWH-18-1-0089). The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, approval, or decision to submit the manuscript.

CONFLICT OF INTEREST STATEMENT

JAG reports grant money to the University of Maryland from the Army Research Institute for the Behavioral and Social Sciences

(W911NF-21-S-0007) to conduct research. The other authors declare no conflicts of interest.

ORCID

Elizabeth D. Rosenman https://orcid.org/0000-0001-9891-3323
Rosemarie Fernandez https://orcid.org/0000-0001-9588-4424

REFERENCES

1. Kozlowski SW, Gully SM, McHugh PP, Salas E, Cannon-Bowers JA. A dynamic theory of leadership and team effectiveness: developmental and task contingent leader roles. *Res Pers Hum Resour Manag*. 1996;14:253-306.
2. Fernandez R, Rosenman ED, Olenick J, et al. Simulation-based team leadership training improves team leadership during actual trauma resuscitations: a randomized controlled trial. *Crit Care Med*. 2020;48(1):73-82. doi:10.1097/CCM.0000000000004077
3. Rosenman ED, Ilgen JS, Shandro JR, Harper AL, Fernandez R. A systematic review of tools used to assess team leadership in health care action teams. *Acad Med*. 2015;90(10):1408-1422. doi:10.1097/ACM.0000000000000848
4. Sebok-Syer SS, Lingard L, Panza M, Van Hooren TA, Rassbach CE. Supportive and collaborative interdependence: distinguishing residents' contributions within health care teams. *Med Educ*. 2023;57(10):medu.15064. doi:10.1111/medu.15064
5. Sebok-Syer SS, Shaw JM, Asghar F, Panza M, Syer MD, Lingard L. A scoping review of approaches for measuring 'interdependent' collaborative performances. *Med Educ*. 2021;55(10):1123-1130. doi:10.1111/medu.14531
6. Griffin DJ, Somaraju AV, Dishop C, DeShon RP. Evaluating interdependence in workgroups: a network-based method. *Organ Res Methods*. 2023;26(3):459-498. doi:10.1177/10944281211068179
7. Hærem T, Pentland BT, Miller KD. Task complexity: extending a core concept. *Acad Manag Rev*. 2015;40(3):446-460. doi:10.5465/amr.2013.0350
8. Leenstra NF, Jung OC, Cnossen F, Jaarsma ADC, Tulleken JE. Development and evaluation of the taxonomy of trauma leadership skills–shortened for observation and reflection in training: a practical tool for observing and reflecting on trauma leadership performance. *Simul Healthc*. 2021;16(1):37-45. doi:10.1097/SIH.0000000000000474
9. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49(2):161-173.
10. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048-1060.
11. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270-292.
12. Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weinger MB. Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. *Simul Healthc*. 2017;12(2):69-75. doi:10.1097/SIH.0000000000000199
13. Debnath SC, Lee BB, Tandon S. Fifty years and going strong: what makes behaviorally anchored rating scales so perennial as an appraisal method? *Int J Bus Soc Sci*. 2015;6(2):57.
14. Rosenman ED, Bullard MJ, Jones KA, et al. Development and empirical testing of a novel team leadership assessment measure: a pilot study using simulated and live patient encounters. *AEM Educ Train*. 2019;3(2):163-171. doi:10.1002/aet2.10321
15. Florez AR, Shepard LN, Frey ME, et al. The concise assessment of leader management tool: evaluation of healthcare provider leadership during real-life pediatric emergencies. *Simul Healthc*. 2023;18(1):24-31.
16. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol*. 1995;50:741-749. doi:10.1037/0003-066X.50.9.741
17. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837. doi:10.1046/j.1365-2923.2003.01594.x
18. Rosenman ED, Branzetti JB, Fernandez R. Assessing team leadership in emergency medicine: the milestones and beyond. *J Grad Med Educ*. 2016;8(3):332-340. doi:10.4300/JGME-D-15-00400.1
19. Leenstra NF, Jung OC, Johnson A, Wendt KW, Tulleken JE. Taxonomy of trauma leadership skills: a framework for leadership training and assessment. *Acad Med*. 2016;91(2):272-281. doi:10.1097/ACM.0000000000000890
20. Rosenman ED, Shandro JR, Ilgen JS, Harper AL, Fernandez R. Leadership training in health care action teams: a systematic review. *Acad Med*. 2014;89(9):1295-1306. doi:10.1097/ACM.0000000000000413
21. Rosenman E, Vrablik M, Brolliar S, Chipman A, Fernandez R. Targeted simulation-based leadership training for trauma team leaders. *West J Emerg Med*. 2019;20(3):520-526. doi:10.5811/westjem.2019.2.41405
22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-429.
23. Fernandez R, Rosenman ED, Brolliar S, et al. An event-based approach to measurement: facilitating observational measurement in highly variable clinical settings. *AEM Educ Train*. 2020;4(2):147-153. doi:10.1002/aet2.10395
24. Committee on Trauma. *Advanced Trauma Life Support for Doctors: Student Course Manual*. 10th ed. American College of Surgeons; 2018.
25. Holcomb JB, Dumire RD, Crommett JW, et al. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma Acute Care Surg*. 2002;52(6):1078-1086.
26. Lubbert PHW, Kaasschieter EG, Hoorntje LE, Leenen LPH. Video registration of trauma team performance in the emergency department: the results of a 2-year analysis in a level 1 trauma center. *J Trauma*. 2009;67(6):1412-1420. doi:10.1097/TA.0b013e31818d0e43
27. Kelleher DC, Bose RJC, Waterhouse LJ, Carter EA, Burd RS. Effect of a checklist on advanced trauma life support workflow deviations during trauma resuscitations without pre-arrival notification. *J Am Coll Surg*. 2014;218(3):459-466.
28. Sugrue M, Seger M, Kerridge R, Sloane D, Deane S. A prospective study of the performance of the trauma team leader. *J Trauma*. 1995;38(1):79-82. doi:10.1097/00005373-199501000-00021
29. Ritchie PD, Cameron PA. An evaluation of trauma team leader performance by video recording. *Aust NZ J Surg*. 1999;69(3):183-186. doi:10.1046/j.1440-1622.1999.01519.x31
30. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420.
31. R Core Team. R: A language and environment for statistical computing. 2023 R Foundation for Statistical Computing, Vienna, Austria.
32. Revelle W. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois; 2017.
33. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*. 1959;56(2):81-105. doi:10.1037/h0046016
34. Widaman KF. Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl Psychol Meas*. 1985;9(1):1-26.
35. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J*. 1999;6(1):1-55.

36. Rosseel Y. Lavaan: an R package for structural equation modeling. *J Stat Softw*. 2012;48(2):1-36. doi:10.18637/jss.v048.i02
37. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008;2(4): 1360-1383. doi:10.1214/08-AOAS191
38. Goodrigh B, Gabry J, Ali I, Brilleman S. Rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.4. 2023.
39. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284-290. doi:10.1037/1040-3590.6.4.284
40. Klein KJ, Ziegert JC, Knight AP, Xiao Y. Dynamic delegation: shared, hierarchical, and deindividualized leadership in extreme action teams. *Adm Sci Q*. 2006;51(4):590-621.
41. Morgeson FP, DeRue DS, Karam EP. Leadership in teams: a functional approach to understanding leadership structures and processes. *J Manag*. 2010;36(1):5-39. doi:10.1177/0149206309347376

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rosenman ED, Grand JA, Fernandez R. Validity evidence of a resuscitation team leadership assessment measure for use in actual trauma resuscitations. *AEM Educ Train*. 2025;9:e11061. doi:10.1002/aet2.11061