The Mayo High Performance Teamwork Scale: Reliability and Validity for Evaluating Key Crew Resource Management Skills

James F. Malec, PhD; Laurence C. Torsher; William F. Dunn; Douglas A. Wiegmann; Jacqueline J. Arnold; Dwight A. Brown; and Vaishali Phatak

Purpose: To develop and evaluate a participant rating scale for assessing high performance teamwork skills in simulation medicine settings.

Methods: In all, 107 participants in crisis resource management (CRM) training in a multidisciplinary medical simulation center generated 273 ratings of key CRM skills after participating in two or three simulation exercises. These data were analyzed using Rasch and traditional psychometric approaches to develop the 16-item Mayo High Performance Teamwork Scale (MHPTS). Sensitivity to change as a result CRM training was also evaluated.

Results: The MHPTS showed satisfactory internal consistency and construct validity by Rasch (person reliability = 0.77; person separation = 1.85; item reliability = 0.96; item separation = 5.04) and traditional psychometric (Cronbach's alpha = 0.85) indicators. The scale demonstrated sensitivity to change as a result of CRM training (pretraining mean = 21.44 versus first posttraining rating mean = 24.37; paired $t = -4.15$, $P < 0.0001$; first posttraining mean = 24.63 versus second posttraining mean = 26.83; paired $t = -4.31$ $P < 0.0001$).

Conclusions: The MHPTS provides a brief, reliable, practical measure of CRM skills that can be used by participants in CRM training to reflect on and evaluate their performance as a team. Further evaluation of validity and appropriateness in other simulation and medical settings is desirable.

(*Simul Healthcare* 2007;2: 4–10)

The principles of training high-performance teams to avoid and manage hazardous errors were developed by the airline industry[1,2] and can be effectively applied in medical settings.[3–6] Contemporary crew or crisis resource management (CRM) training includes instruction and practice to develop situation awareness, communication skills, anticipation of error chains, as well as error containment and management strategies.[7,8] Curricula for training CRM behavioral skills are essential to widespread implementation of CRM in actual medical settings and have become a key component of training in medical simulation centers.[9,10]

There is general agreement that behavioral rating methods are most appropriate to assess training in these nontechnical skills[9,11,12] although standardized and widely accepted methods for measuring the outcomes of CRM training in medicine are not presently available. Fletcher and associates[12] identified four categories of behaviors for training and evaluation in CRM: 1) *cooperation/communication*, such as team building and maintenance, consideration and support of others, and conflict solving; 2) *leadership/management*, including use of authority and assertiveness, providing and maintaining standards, planning, coordination, and workload management; 3) *situation awareness*, such as system awareness, environmental awareness, and awareness of time and anticipation of future events; and 4) *decision making*, including problem definition/diagnosis, option generation, risk assessment, option selection, and outcome review. Based on these categories of nontechnical skills, Fletcher's group developed the Anesthetists' Nontechnical Skills (ANTS) evaluation method to rate individual team member performance using clearly specified behavioral markers.[12] The ANTS has satisfactory accuracy, internal consistency, and usability with acceptable levels of interrater agreement. Kim and colleagues[13] have reported satisfactory construct validity, internal consistency, and interrater reliability for a measure of team functioning in CRM settings, the Ottawa CRM Global Rating Scale (Ottawa GRS). The Ottawa GRS uses global rating scales of eight important aspects of medical teamwork—problem solving, situational awareness, anticipation and planning, leadership, resource utilization, communication output, communication input, and clinical management—as well as an overall rating of team performance. Weller and colleagues[14] reported good expert rater agreement for global ratings of performance in simulated medical crisis management. Most appropriately used by expert raters, these scales' statistical psychometric properties have not been definitively established. Additional research is required is to examine

From the Departments of Psychiatry and Psychology (J.F.M., V.P.), Anesthesiology (L.C.T.), Pulmonary and Critical Care Medicine (W.F.D.), Cardiovascular Surgery (D.A.W.), and Nursing (J.J.A.), and the Mayo Quality Office (D.A.B.), Mayo Multidisciplinary Simulation Center and Mayo Clinic College of Medicine, Rochester, MN.

The authors have indicated they have no conflict of interest to disclose.

Reprints: James F. Malec, PhD, Tertiary Psychiatry and Psychology, Generose ME, Mayo Clinic, Rochester, MN 55905 (e-mail: malec.james@mayo.edu).

ISSN: 1559-2332/07/0201-0004

features such as dimensionality, item separation, and validity for each of these scales.

In this study, we developed and conducted initial psychometric evaluation of a behavioral rating scale to assess CRM skills. In designing the original version of the measure, we referenced behaviors identified by Fletcher and colleagues as well as behaviors targeted for training in CRM curricula as recommended by Gaba[9] and Raemer.[15] Primary objectives of this project were to develop a scale that is brief for practical use in training settings, evaluates behaviors of a medical team that represent a range of CRM skills, and is sufficiently behavioral and transparent to be used reliably even by naive training participants with little or no CRM knowledge or experience. A scale with these features may be of value not only to assess the effectiveness of training but also to engage learners in a self-reflective educational process aimed at improving their awareness of and skill in high performance team processes. A primary goal was to develop a reliable scale to guide self-reflection and self-assessment to facilitate debriefing.

To develop such a scale, we employed both traditional and more contemporary measurement techniques, specifically Rasch analysis.[16,17] Rasch analysis provides information about the scaling and the unique contribution of individual items. This type of analysis is helpful in selecting and developing optimal scaling for a small set of items that nonetheless represents the spectrum of behaviors that describes the underlying construct of interest, in this case, CRM skills. Although we present initial information about the variability and sensitivity of this measure over the course of training, the focus of this study was to identify and scale a set of items that represented a range of CRM skills. Rasch analysis assumes a one-dimensional underlying structure to the construct of interest and selects items that fit with this single dimension. Identifying key items that represent the range of a skill precludes the necessity of exhaustively and comprehensively describing all aspects of a skill. In this way, Rasch analysis supports the development of brief measures that nonetheless possess satisfactory reliability and validity. This psychometric method is being increasingly used for development of measures relevant to healthcare outcomes[18–20] and provides a basis for innovative measurement procedures such as computerized adaptive testing.[19]

METHOD

Subjects

Participants were 19 residents and 88 nurses who participated in CRM training in the Mayo Multidisciplinary Simulation Center in Rochester, MN.

Measure

The original version of the measure consisted of 19 items describing behaviors believed to characterize high performance teamwork. These behaviors were the focus of the CRM training. All ratings were completed by participants through reflective analysis of the performance of their team in the CRM training in which they were participating. In the original measure, each item was rated on a four-point scale indicating how often the behavior occurred in the simulation scenario: 1 = never; 2 = rarely; 3 = inconsistently; 4 = consistently. For all items but one, a higher score indicated better performance. For one item that was ultimately eliminated from the scale ("The team becomes fixated on an isolated indicator or occurrence to the exclusion of other important aspects of care"), scoring was reversed. For some items that were not required in all situations (items 9–16, Table 1), participants had the option of indicating "not applicable."

Procedure

During the CRM training sessions, participants were engaged in a simulation scenario following a brief orientation but prior to extensive CRM didactic education or practice. Following their participation in this first scenario, participants retrospectively rated the performance of their team using the

TABLE 1. Mayo High Performance Teamwork Scale

Use the following scale to rate the team on each dimension:

0	**1**	**2**
Never or rarely	**Inconsistently**	**Consistently**

Please rate conservatively. Most teams that have not worked extensively together do not consistently demonstrate many of the qualities described in the scale.

Always rate items 1–8.

_____ (1) A leader is clearly recognized by all team members.

_____ (2) The team leader assures maintenance of an appropriate balance between command authority and team member participation.

_____ (3) Each team member demonstrates a clear understanding of his or her role.

_____ (4) The team prompts each other to attend to all significant clinical indicators throughout the procedure/intervention.

_____ (5) When team members are actively involved with the patient, they verbalize their activities aloud.

_____ (6) Team members repeat back or paraphrase instructions and clarifications to indicate that they heard them correctly.

_____ (7) Team members refer to established protocols and checklists for the procedure/intervention.

_____ (8) All members of the team are appropriately involved and participate in the activity.

Items 9–16 may be marked "NA (not applicable)" if no situations occurred in which these types of responses were required.

_____ (9) Disagreements or conflicts among team members are addressed without a loss of situation awareness.

_____ (10) When appropriate, roles are shifted to address urgent or emergent events.

_____ (11) When directions are unclear, team members acknowledge their lack of understanding and ask for repetition and clarification.

_____ (12) Team members acknowledge—in a positive manner—statements directed at avoiding or containing errors or seeking clarification.

_____ (13) Team members call attention to actions that they feel could cause errors or complications.

_____ (14) Team members respond to potential errors or complications with procedures that avoid the error or complication.

_____ (15) When statements directed at avoiding or containing errors or complications do not elicit a response to avoid or contain the error, team members persist in seeking a response.

_____ (16) Team members ask each other for assistance prior to or during periods of task overload.

original measure. Subsequently, they debriefed on the scenario with the instructor and received specific instruction and feedback about CRM and teamwork. Then they participated in another scenario and rated their team's performance in this second scenario on the measure followed by a debriefing session. In some cases, time was available after debriefing the second scenario for participation, team rating, and debriefing for a third scenario.

Scenarios were either designed to represent critical anesthesia management (AM) or emergency response team (ERT) situations. Ratings were done from memory immediately after the completion of each scenario. All participants were actively involved in all ERT scenarios. In the AM scenarios, some participants only observed during one of the scenarios. The AM scenarios varied among different learner groups. The ERT scenarios were consistent across groups; all involved response to various types of cardiac arrhythmias. The first scenario involved response to a fixation error; the second to distraction; the third involved no planned negative event. Hence, the measurement tool was applied to a variety of scenarios to minimize the potential of bias resulting from restricting its application to only one or two scenarios.

Analyses

Rasch analysis and traditional psychometric approaches were used to develop and evaluate the Mayo High Performance Team Scale (MHPTS; Table 1). Rasch analysis is a method to develop mathematical models for rating scale data based on the hierarchical ordinal relationship among rated items. In a rating scale that accurately and reliably represents the construct being measured, items should have a consistent and predictable hierarchical relationship to each other. Independent of raters or the events being rated, some items should consistently be rated as occurring frequently at a high level ("easier" items) while other items should reliably be observed less frequently and at lower levels of performance ("harder" items). A range of such items should consistently categorize people as high, low, or medium performers. That is, those with high scores on the "harder" items should also get high scores on the "easier" items and on the scale overall, and vice versa. Person and Item Reliability and Separation are metrics that indicate the strength of these relationships and, as such, the integrity of the scale. The degree to which individual items are consistent with the underlying model are indicated by Fit metrics.

In this study, all participants rated a small number of distinct events, that is, performance of the team in simulated medical/surgical scenarios. Metrics for Person Reliability and Separation essentially represented expected rater bias and reflected the consistency of their biases across items. It is desirable for items to be equally sensitive to rater bias. That is, if a rater tends to rate conservatively, the rater's conservative bias would ideally be consistent across all items. Of most interest to this analysis were metrics for Item Fit, Reliability, and Separation. These metrics represented the consistency of raters' perceptions of the difficulty of each item in relation to other items.

In traditional psychometrics, Cronbach's alpha provides an indication of the internal consistency of a scale. Item-to-scale correlations provide a measure of the degree to which individual items "fit" with the entire scale. Although based on different mathematical models, both Rasch and traditional psychometric analyses are expected to yield complementary positive results for reliable and valid measures. Brief definitions of metrics and statistics used in this study are provided in Table 2.

RESULTS

Rasch Item Analyses

Rasch analysis was applied to the data for 273 separate performance assessments using the 19 original rating items. In all, 106 of these assessments were obtained prior to training, 107 of a second scenario following training, and 60 after a third scenario subsequent to training. For the initial analyses of item metrics and fit, we used the entire sample of ratings. Some of these ratings were made by the same person at various points in the training sequence and thus provided examples of ratings made at increasing levels of experience with CRM and the rating scale.

Examination of item metrics showed that the two lowest ends of the rating scale ("never" and "rarely") were not clearly distinct. For most items, the lowest end of the rating scale was used in less than 5% of the ratings. Therefore, we

TABLE 2. Psychometric Glossary

Reliability: Consistency of a measure; degree to which the measure yields the same score when measuring the same object or event.

Validity: Accuracy of a measure; degree to which the measure accurately represents the construct that it purports to measure.

Person Reliability: The consistency across items of raters giving lower scores to more difficult items and higher scores to easier items. Range = 0 to 1.0. Target value: 0.80 or greater.

Person Separation: Indicator of power of items to distinguish among raters. Target value: 2.0 or greater.

Item Reliability: The consistency across raters of the hierarchical structure of the items, i.e., ranking particular items as more difficult than others. Range = 0 to 1.0. Target value: 0.90 or greater.

Item Separation: Degree to which items represent distinct aspects of the measurement scale. Target value: 4.0 or greater.

Item Infit: Degree to which an item maintains its position in the hierarchy of items. Very low values indicate overfitting or redundancy. High values indicate that an item may not consistently represent the construct being measured. Target value: 0.6 to 1.4.

Item Outfit: Degree to which an item represents extreme cases. Target value: 0.6 to 1.4.

Cronbach's alpha: Internal consistency of a measure; degree to which subsets of items represent the entire scale. Range = 0 to 1.0. Target value: 0.80 or greater.

Pearson correlation: Relationship of two sets of test scores. Range = −1.0 to +1.0. Correlations of +1.0 indicate that high scores on one measure perfectly predict high scores on another measure; correlations of −1.0 indicate that high scores on one measure perfectly represent low scores on the other measure.

Item-to-scale point-biserial correlation: Correlation of individual item scores with the total score for the measure. Target value: +0.40 to +0.80.

Definitions and target values are specific to this study to a degree; interested readers are referred to original sources (eg, Bond and Fox[17]; Wright and Masters[16]; Anastasi[22]) for more detailed and mathematical descriptions of these terms.

combined the two lowest levels resulting in a three-level scale for the revised measure: never/rarely = 0; inconsistently = 1; consistently = 2. Several items (Table 1; items 9–16) were not always rated because the events to be rated did not occur in all simulations. We reasoned that the absence of a negative event in these cases was more positive than a poor response. Therefore, for these items, we scored the absence of a rating as 1, whereas for items 1–8, the absence of a rating was scored 0. Further examination of items 9 to 16 suggested that they performed differently than items 1 to 8. Items 9–16 identify adaptive behaviors in response to a potentially negative event. A poor response to such events typically increases the negative impact of the event and, in the recommended scaling system, results in a score of "0" for the item, ie, loss of the point given for that item. A mediocre response results in no gain, ie, score for the item remains "1." An adaptive response results in an additional point for the item with a score of "2." Recoding the original 19 item scale in this way resulted in improved fit for items 9–16 and in an overall improvement in Person Reliability from 0.71 to 0.79 with negligible change in Item Reliability from 0.98 to 0.97.

The next steps in this analysis were conducted using this revised rating and scoring system to identify items which best represented the construct of high performance teamwork. Rasch analysis of the original 19 items with the WINSTEPS[21] program revealed lack of fit for the item "The team becomes fixated on an isolated indicator or occurrence to the exclusion of other important aspects of care" (Item 17, Table 3; Infit Mean Square (MNSQ) = 1.60; Outfit MNSQ = 1.63). This item may have been overly specific. This same analysis indicated overfitting (ie, redundancy) for the item, "When statements directed at avoiding or containing errors or seeking clarification are not acknowledged, team members persist in seeking acknowledgment" (Item 18, Table 3; Infit MNSQ = 0.66; Outfit MNSQ = 0.70). This item was frequently not rated. Furthermore, this item (acknowledging a challenge) is a necessary part and consequently partially redundant with item 15 (Table 1). A subsequent Rasch analysis was run eliminating these two items. In this analysis, the item, "Material resources are used cost effectively throughout the activity," again showed inadequate fit as it had in the initial analysis (Item 18, Table 3; Infit MNSQ = 1.44; MNSQ Outfit = 1.31). This item may have been overly general and vague and was eliminated.

Rasch analysis of the remaining 16 items showed satisfactory Infit and Outfit for all items; item-to-scale point-biserial correlations ranged from0.37 to 0.64 (Table 3). Item 9 fit least well. Item 9 was often left unrated with 41% missing values. However, examination of the Rasch item difficulty level (Table 4) suggested that, for most raters, this item represented a particularly stringent test of teamwork relative to the other items. For this reason, item 9 was retained in the revised scale despite its marginal fit.

The difficulty level of each item represents the frequency of high versus low scores on the item, that is, more difficult items get higher ratings less frequently than less difficult items. Difficulty levels for both the original and revised scales are provided in Table 4. Higher positive scores indicate more difficult items, that is, items that raters felt were more stringent and rigorous indications of a high level of team performance. Lower negative scores indicate easier items, that is, team behaviors that are more commonplace and represent a lower level of team performance. Overall raters considered items 9 and 15 (Table 4) to represent the most

TABLE 3. Infit and Outfit Mean Squares (MNSQ) and Item-scale Correlations for Items in Original and Revised Measure

	Original Measure			Revised Measure		
Item	Infit MNSQ	Outfit MNSQ	Item-scale Correlation	Infit MNSQ	Outfit MNSQ	Item-scale Correlation
17	1.60	1.63	0.28	—	—	—
19	1.44	1.31	0.52	—	—	—
9	1.29	1.36	0.32	1.32	1.42	0.37
1	1.23	1.06	0.60	1.26	1.11	0.59
7	1.23	1.18	0.59	1.28	1.23	0.58
8	1.15	0.80	0.54	1.20	0.89	0.51
6	1.06	0.98	0.63	1.11	1.02	0.62
5	1.02	0.95	0.52	1.07	1.02	0.50
10	1.00	1.03	0.48	1.03	1.10	0.49
2	0.98	0.84	0.65	1.00	0.86	0.64
16	0.97	1.02	0.45	0.99	0.99	0.48
12	0.95	0.96	0.50	0.97	1.09	0.50
4	0.91	0.82	0.55	0.92	0.94	0.53
3	0.90	0.81	0.65	0.92	0.82	0.63
14	0.85	0.89	0.45	0.90	0.93	0.46
11	0.76	0.75	0.59	0.78	0.81	0.60
13	0.73	0.74	0.51	0.77	0.82	0.52
15	0.69	0.74	0.49	0.75	0.77	0.50
18	0.66	0.70	0.48	—	—	—

TABLE 4. Item Difficulty for Original and Revised Scale Including All Observations and for Revised Scale at Each Assessment Time

Item	Original	Revised	Pretraining	Posttraining 1	Posttraining 2
17	1.64	—	—	—	—
9	0.85	1.01	0.87 (16)	0.70 (14)	2.42 (16)
15	0.74	0.90	0.71 (14)	0.76 (16)	2.00 (15)
18	0.70	—	—	—	—
11	0.49	0.63	0.52 (13)	0.67 (13)	1.05 (13)
13	0.44	0.58	0.30 (10)	0.76 (16)	1.13 (14)
6	0.35	0.48	0.73 (15)	0.33 (10)	0.31 (9)
14	0.19	0.31	−0.02 (7)	0.49 (12)	0.98 (12)
7	0.12	0.24	0.33 (11)	0.43 (11)	−0.53 (8)
16	−0.12	−0.02	−0.39 (5)	0.20 (9)	0.66 (10)
10	−0.16	−0.06	−0.26 (6)	−0.16 (8)	0.90 (11)
2	−0.20	−0.10	0.27 (9)	−0.24 (7)	−1.12 (4)
1	−0.23	−0.13	0.33 (11)	−0.44 (4)	−0.95 (5)
3	−0.40	−0.30	0.01 (8)	−0.35 (6)	−1.52 (2)
19	−0.56	—	—	—	—
5	−0.64	−0.56	−0.52 (4)	−0.61 (3)	−0.41 (7)
12	−0.77	−0.69	−0.81 (2)	−0.44 (4)	−0.95 (5)
4	−0.91	−0.84	−0.81 (2)	−0.66 (2)	−1.52 (2)
8	−1.50	−1.44	−1.26 (1)	−1.45 (1)	−2.44 (1)

stringent tests of teamwork skills and items 4 and 8 to represent easier, more commonly demonstrated skills.

For the final 16-item Mayo High Performance Teamwork Scale (MHPTS; Table 1), Person Reliability = 0.77; Person Separation = 1.85; Item reliability = 0.96; Item Separation = 5.04. Person reliability indicators satisfactorily approximated target values considering the measurement context and naïveté of the raters. Item reliability indicators were above target values.

Recognizing that small samples at each assessment point challenge the reliability of statistics computed separately for each point, the properties of the scale were examined separately for each of the pretraining and two posttraining assessments (Table 4). To facilitate comparison in Table 4, the ordinal rank of each item on difficulty level is also provided for each of the three assessment times. Items generally hold their rank within high, low, and mid ranges for the scale across assessment times, particularly for extreme items. Item Reliability remained acceptable at each assessment point: pretraining = 0.91; posttraining 1 = 0.89; posttraining 2 = 0.91, despite the relatively small samples at each time. Person Reliability was also relatively consistent across the pretraining and first posttraining assessments but declined in the very small sample with data at the second postassessment: pretraining = 0.79; posttraining 1 = 0.75; posttraining 2 = 0.67.

Internal Consistency/Construct Validity

Rasch analysis provides an interval equivalent metric for the derived scale. The sample used in this study was not large and may not be representative of the population of raters, however, to support definitive calibration of the MHPTS. Furthermore, conversion of raw scores to Rasch calibrated scores may be cumbersome in many training settings. The Pearson correlation between the Rasch-derived metric and the raw score for the MHPTS was 0.96, indicating that the two values were essentially equivalent and that a simple additive model provides a reasonable approximation of the Rasch model. For ease of calculation of the summary score, particularly in other settings, we used the raw score in subsequent analyses. The distribution of raw scores reveals a positive skew (Fig. 1), that is, raters tended to give high scores.

Item and Person Reliability statistics provide indicators of internal consistency and fit with the putative underlying construct. We also computed the more traditional measure of internal consistency, Cronbach's alpha. For all ratings

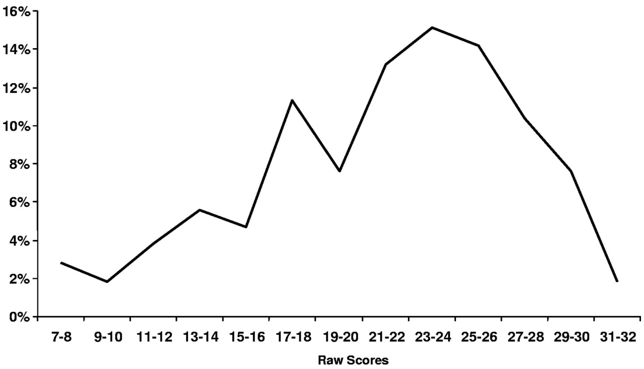

FIGURE 1. Distribution of total raw scores for the pretraining MHPTS assessments. Total raw score is on the abscissa; percent of raters giving that score is on the ordinal. A positive skew of the distribution is apparent.

Cronbach's alpha = 0.85. For the 106, pretraining ratings, Cronbach's alpha = 0.83. For 107 ratings made in the first posttraining scenario, Cronbach's alpha = 0.83. For 60 ratings of a second posttraining scenario, Cronbach's alpha = 0.81.

Validity as a Measure of Change with Training

To evaluate the sensitivity of the MHPTS to effects of CRM training, we conducted paired *t* tests comparing raw MHPTS scores for the pretraining scenario to those for the first posttraining scenario and, where available, for the first and second posttraining scenarios. Analysis of the 106 assessments by the same rater for both the pretraining and the first posttraining scenarios showed a statistically significant improvement (pretraining mean = 21.44; posttraining mean = 24.37; paired $t = -4.15$ $P < 0.0001$). The mean difference was −2.93 (SD of the difference = 7.25; standard error of the mean difference = 0.71).

Analysis of the smaller sample (n=59) who had rated the pretraining scenario as well as two posttraining scenarios also revealed a significant improvement between the pretraining assessment (mean = 21.83) and the first posttraining assessment (mean = 24.63; paired $t = -3.38$, $P < 0.001$) as well as between assessments of the first and second posttraining scenarios (first post-training mean = 24.63; second post-training mean = 26.83; paired $t = -3.93$ $P < 0.0001$). The mean difference between the pretraining and first post-training assessments was −2.80 (SD of the difference = 6.36; standard error of the mean difference = 0.83). The mean difference between the first and second posttraining assessments was −2.20 (SD of the difference = 4.31; standard error of the mean difference = 0.56). Figure 2 displays these changes. The subsample who rated a third scenario tracked very closely with the full sample for the pretraining and first posttraining assessments, suggesting that the rating behavior of these two rater groups was very similar.

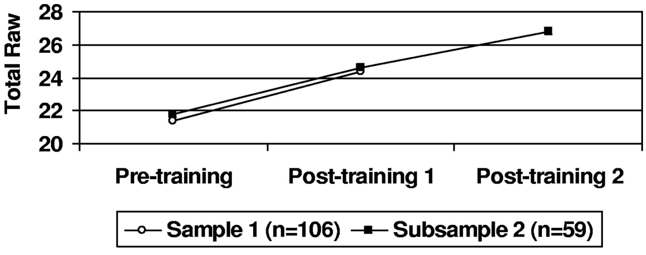

FIGURE 2. Significant improvement in CRM skills across 106 asessements conducted at pretraining debriefing (mean = 21.4) and the first posttraining debriefing (mean = 24.4; paired $t = -4.15$, $P < 0.0001$). A highly significant improvement is also apparent for a smaller sample of 59 who conducted three assessments. Improvement occurred between the pretraining rating (mean = 21.8) and the first posttraining rating (mean = 24.6; paired $t = -3.38$, $P < 0.001$) and between ratings of the first and second posttraining scenarios (first posttraining mean = 24.6; second posttraining mean = 26.8; paired $t = -3.93$ $P < 0.0001$).

DISCUSSION

The Mayo High Performance Teamwork Scale (MHPTS) was designed to be sufficiently brief, behavioral, and understandable to be used practically by naive participants in training and other settings to rate key behaviors of high performance teams. Rasch analyses reported here were used to refine the MHPTS. The resulting measure possesses satisfactory reliability when used by naive raters as indicated both by Rasch and traditional psychometric indices. Although not exhaustive in describing all possible behaviors that characterize high performance teams, the MHPTS items provide a representative sample of the range of key behaviors of such teams. This method of scale development resulted in a relatively brief measure that can be practically used in training settings.

Results reported here demonstrate the construct validity of the MHPTS and that it is sensitive to change as a result of training in a CRM course. Establishing the validity of behavioral measures, however, requires extensive testing, particularly when a "gold standard" is not available for comparison. Future studies that use the MHPTS in comparison to other types of scales measuring the same domain will be helpful in further validating these measures and assessing the strengths, weaknesses and most appropriate applications of each approach. The comparison of self-ratings with expert assessments will be of particular interest to cross-validate these measures and investigate biases inherent in each type of assessment. Contrasting trainee and expert assessments over the course of training will be important to study the process of change in trainee self-assessments.

Further validation and refinement of this measure will also require its evaluation with larger samples in other centers and settings using a variety of team-oriented medical simulation scenarios. Additional data and analyses of this type are necessary to increase confidence that the measure truly represents the construct of medical teamwork for the population of potential trainees within the universe of possible team-oriented scenarios. Larger data sets will allow more detailed examination of the integrity of individual items and of differential item functioning. Initial analyses reported here suggest that the item structure of the measure is grossly stable over the course of a single training session. However, variability in the relationships among items was also apparent (Table 4). To determine whether this variability represents random or systematic variation will require additional study with larger and more diverse samples.

In the study reported here, raters tended to give high scores. This may be an inherent bias in naive raters and groups rating their own performance. In an attempt to temper this bias, an instruction was added to the revised scale to "Please rate conservatively. Most teams that have not worked extensively together do not consistently demonstrate many of the qualities described in the scale." From one perspective, the use of naive raters is a limitation of the study. Reliability and validity would be expected to improve on any measure by using well-trained, expert raters. However, an important objective of the study was to design a measure that could be used for self-evaluation and self-reflection by learners in

CRM training settings. Results suggest that the MHPTS can be used with reasonable reliability even by naive raters. The inherent bias in self-ratings must also be recognized in using the MHPTS as a measure of the efficacy of CRM training. In evaluating the efficacy of CRM programs, expert ratings will also be of value. Recognizing potential bias among both expert and relatively naive raters, ratings of training programs from multiple perspectives may be optimal.

In summary, the MHPTS provides a brief measure of a range of high performance teamwork skills that are the target of CRM training in medical settings. The MHPTS appears sufficiently behavioral and transparent to be used reliably even by naive training participants with little or no CRM knowledge or experience. Results reported here provide evidence of satisfactory reliability and initial support for the construct validity and sensitivity to change of the measure. Further evaluation of the MHPTS in other simulation settings as well as actual medical settings is necessary to further assess its validity in various educational and clinical settings.

ACKNOWLEDGMENTS

This study was supported in part by an Education Innovation grant from the Mayo Foundation in Rochester, MN.

REFERENCES

1. Wiener EL, Kanki BG, Helmreich RL. *Cockpit resource management.* San Diego, CA: Academic Press; 1993.
2. Cooper GE, White MD, Lauber JK. Resource management on the flightdeck: proceedings of a NASA/Industry Workshop. Moffett Field, CA: NASA-Ames Research Center; 1980.
3. Shortell SM, Zimmerman JE, Rousseau DM, et al. The performance of intensive care units: does good management make a difference? *Medical Care* 1994; 32:508–525.
4. Barach P, Small SD: Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *Br Med J* 2000; 320: 759–763.
5. Howard SK, Gaba DM, Fish KJ, et al: Anesthesia crisis resource management training: teaching anesthesiologist to handle critical incidents. *Aviation Space Envir Med* 1992; 63:763–770.
6. Grogan EL, Stiles RA, France DJ, et al: The impact of aviation-based teamwork training on the attitudes of health-care professionals. *J Am College Surg* 2004; 199:843–848.
7. Helmreich RL, Wilhelm JA, Klinect JR, Merritt AC: Culture, error, and crew resource management. In: Salas E, Bowers CA, Edens E, eds. *Improving teamwork in organizations.* Hillsdale, NJ: Erlbaum; 2001: 305–331.
8. Reason J.*Human error.* New York: Cambridge University; 1990.
9. Gaba DM, Howard SK, Fish KJ, et al: Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. *Simul Gaming* 2001; 32:175–193.
10. Dunn WF, ed. *Simulators in critical care and beyond.* Des Plaines, IL: Society for Critical Care Medicine; 2004.
11. Wright MC, Taekman JM, Endsley MR: Objective measures of situation awareness in a simulated medical environment. *Qual Safety Health Care.* 2004; 13 (Suppl 1):65–71.
12. Fletcher G, Flin R, McGeorge P, et al: Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioral marker system. *Br J Anaesth* 2003; 90:580–588.
13. Kim J, Neilipovitz D, Cardinal P, et al: 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; 34:2167–2174.
14. Weller JM, Bloch M, Young S, et al: Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *Br J Anaesth* 2003; 90:43–47.
15. Raemer D. Team-oriented medical simulation. In: Dunn WF, ed. *Simulator in Critical Care and Beyond.* Des Plaines, IL: *Society of Critical Care Medicine*; 2004:42–46.
16. Wright BD, Masters GN. *Rating scale analysis: Rasch measurement.* Chicago: MESA Press; 1982.
17. Bond TM, Fox CG. *Applying the Rasch model: fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum; 2001.
18. Malec JF, Kragness M, Evans RW, et al: Further psychometric evaluation and revision of the Mayo-Portland Adaptability Inventory in a national sample. *J Head Trauma Rehab* 2003; 8:479–492.
19. Haley SM, Coster WJ, Andres PL, et al: Score comparability of short forms and computerized adaptive testing: simulation study with the activity measure for post-acute care. *Arch Phys Med Rehab* 2004; 85:661–666.
20. Linn RT, Blair RS, Granger CV, et al: Does the functional assessment measure (FAM) extend the functional independence measure (FIM) instrument? A rasch analysis of stroke inpatients. *J Outcome Meas* 1999; 3:339–359.
21. *WINSTEPS: Rasch-Model Computer Program* [computer program]. Version 3.32. Chicago, IL: Winsteps; 2001.
22. Anastasi A. *Psychological testing.* New York: McMillan; 1988.