Measuring Ward-Based Multidisciplinary Healthcare Team Functioning: A Validation Study of the Team Functioning Assessment Tool (TFAT)

Gigi Sutton, Jenny Liao, Nerina L. Jimmieson, Simon L. D. Restubog

Abstract: The team functioning assessment tool (TFAT) has been shown to be a reliable behavioral marker tool for assessing nontechnical skills that are critical to the success of ward-based healthcare teams. This paper aims to refine and shorten the length of the TFAT to improve usability, and establish its reliability and construct validity. Psychometric testing based on 110 multidisciplinary healthcare teams demonstrated that the TFAT is a reliable and valid tool for measuring team members' nontechnical skills in regards to Clinical Planning, Executive Tasks, and Team Functioning. Providing support for concurrent validity, high TFAT ratings were predicted by low levels of organizational constraints and high levels of group potency. There was also partial support for the negative relationships between time pressure, leadership ambiguity, and TFAT ratings. The paper provides a discussion on the applicability of the tool for assessing multidisciplinary healthcare team functioning in the context of improving team effectiveness and patient safety for ward-based hospital teams.

Keywords
behavioral marker methodology
crew/crisis resource management
multidisciplinary healthcare teams
nontechnical skills
team functioning

Healthcare organizations have begun to challenge traditional hierarchical structures that promote individualized decision making and, instead, have shifted toward a *collaborative* approach to patient care (Carter, Garside, & Black, 2003). Patient safety can be best improved through enactment of cultural change at the level of the team rather than the organization as a whole (Firth-Cozens, 2001), thereby reflecting an increasing demand for the modern healthcare professional to develop a new set of nontechnical skills relating to team processes and functioning. As such, the assessment of nontechnical skills relating to healthcare team functioning plays a critical first step toward reducing errors in patient care delivery, and promoting effective team outcomes in the healthcare sector.

TFAT Background and Development

The validation of competencies for effective teamwork has been the focus of team research for the last 30 years. Crisis resource management (CRM) has been viewed as a method of improving communication and addressing hierarchical dominated problem solving through team centered decision making, thereby promoting healthier patient outcomes (Kosnik, 2002; Miller, 2005). In order to develop a structured observation list of nontechnical skills within crisis healthcare teams, researchers have employed domain-specific behavioral marker systems (e.g., Yule, Flin, Paterson-Brown, & Maran, 2004). Effectiveness of nontechnical skills within a team setting necessarily reflects the functioning of a team as a whole, rather than the performance of individuals, due to its collaborative nature, sharing of goals, and degree of acceptance of shared membership. However, the majority of work in this area has been developed to assess the behaviors of particular individuals within a team (e.g., anesthetists in surgical teams; Fletcher, et al., 2003) as opposed to an entire team. Moreover, it is has been restricted to use with "crisis" teams (Flin & Maran, 2004; Wilson, Burke, Priest, & Salas, 2005).

The team functioning assessment tool (TFAT) is one tool that has been developed to assess the nontechnical skills of multidisciplinary teams within a ward-based hospital environment (Sutton, Liao, Jimmieson, & Restubog, 2011). The TFAT is based on a taxonomy of behaviors depicting the characteristics of high-performing ward-based multidisciplinary healthcare teams (MDHT). As documented in Sutton and colleagues, a programmatic series of three studies was implemented to develop the taxonomy of behaviors, yielding a preliminary taxonomy of 61 behavioral items. Results provided a preliminary indication of high agreement among observers on TFAT ratings. Moreover, the TFAT was shown to be reliable in regards to inter-rater agreement on TFAT ratings (r_{wg} ranged from .67 to .99) and ranking of three teams using the TFAT (concordance W coefficients ranged from 0.46 to 0.78).

Journal for Healthcare Quality
Vol. 00, No. 0, pp. 1–13

The TFAT isolates critical skills and behaviors for effective team performance (Fletcher et al., 2003; Sutton et al., 2011), and establishes a common language for defining team effectiveness. Intended to be used by trained observers for evaluating the nontechnical skills of MDHTs, the TFAT is designed to evaluate team functioning as defined by three core categories of Clinical Planning, Executive Tasks, and Team Relations. *Clinical Planning* refers to skills and behaviors for systematically collecting, verifying, and interpreting data from multiple sources for generating, implementing, evaluating, and reviewing patient care plans, while also maintaining awareness of the work and patient setting. The *Executive Tasks* category assesses skills and behaviors for organizing resources and activities to achieve identified goals and encourage innovation, while *Team Relations* encompasses the skills and behaviors required for achieving a balance of enquiry and assertion in the relations and communication among team members to promote effective team decision making. The 61 behavioral items in the TFAT have been shown to be reliable based on trained observer ratings. However, in line with Hinkin's (1998) outline of scale development, it is noted that items and categories in the TFAT require construct validation. In addition, a shorter version of the TFAT would be more practical in the applied setting. Our broad program of research sets out to apply the knowledge gained by high-reliability teams to develop methodologies targeted toward measuring *team* functioning within "*non-crisis*" *ward-based* MDHTs with the following aims:

1. Refine and reduce the number of items in the TFAT to improve usability of the tool.
2. Establish the internal consistency of the categories in the revised TFAT.
3. Demonstrate the inter-rater reliability of the revised TFAT.
4. Further establish content validity of items through consultations with subject-matter experts (SMEs).
5. Establish construct validity of the revised TFAT through three means:

(a) Exploratory factor analyses (EFAs) for each of the three core TFAT categories to determine factor structure.
(b) Confirmatory factor analyses (CFAs) to confirm the factor structure of the predicted measurement models in each of the subcategories, as well as to establish the discriminant validity between the subcategories within each of the three core TFAT categories.
(c) Concurrent validity by testing the linkages between TFAT scores and theoretically related variables as perceived by teams.

Linking the TFAT to Job Demand and Job Resource Variables

In order to demonstrate that the shortened version of the TFAT is a valid tool for discriminating between high and low functioning teams, hypotheses linking the TFAT to theoretically relevant constructs were developed. Drawing from the Job Demands-Resources (JD-R) model (Bakker & Demerouti, 2007; Demerouti, Bakker, Nachreiner, & Schaufeli, 2001), variables relating to job demands (i.e., organizational constraints, time pressure, and leadership ambiguity) and job resources (i.e., group potency) should predict team functioning outcomes as measured by the TFAT. According to the JD-R model, job demands are physical, psychological, social, or organizational aspects of a job that require sustained effort, and consequently, can reduce team coordination practices because they are initiators of the stress process. On the other hand, job resources include physical, psychological, social, or organizational aspects of a job that promote performance and learning because they are initiators of motivation.

Organizational constraints were expected to be negatively associated with TFAT ratings, as a lack of instrumental support and resources prevents employees from achieving effective coordination behaviors and high levels of work performance (Spector & Jex, 1998). Similarly, *time pressure* should be negatively linked to TFAT ratings as the experience of strain associated with time pressure has been shown to reduce team functioning (Kelly & Loving, 2004). Lack of clarity over leadership (or *leadership ambiguity*) also has been shown to reduce team coordination (West et al., 2003), and as such, it is predicted that leadership ambiguity is negatively related to TFAT ratings. Given that job resources enable teams to perform better, it also is expected that teams with high levels of *team potency* should be positively linked to TFAT ratings. Indeed, team confidence in capability promotes teams to achieve effective team

functioning outcomes (Guzzo, Yost, Campbell, & Shea, 1993). In sum, the following hypotheses are offered with respect to all three TFAT categories:

Hypothesis 1. Teams reporting high levels of organizational constraints will have lower TFAT ratings.

Hypothesis 2. Teams reporting high levels of time pressure will have lower TFAT ratings.

Hypothesis 3. Teams reporting high levels of leadership ambiguity will have lower TFAT ratings.

Hypothesis 4. Teams reporting high levels of group potency will have higher TFAT ratings.

Method

Participants and Procedure

This research was conducted as part of a larger program of research into team effectiveness which involved a comprehensive approach to ethical approval described in full elsewhere (Sutton, et al., 2011). As part of this process, written informed consent was obtained at the following levels: district head, hospital administrator (CEO/medical superintendent/executive director), hospital divisional medical and nursing directors, departmental allied health directors; and all members of participating MDHTs. Two teams approached were excluded due to a team member declining to participate in the research.

Observational ratings of team functioning were collected from 133 MDHTs across 10 hospital sites in southeast Queensland, Australia. Teams were included in the sample if: (1) team members attended clinical meetings on a regular basis (i.e., at least once per week), (2) team members from at least three professions attended the team meeting, and (3) teams primarily offered an acute, ward-based service. Upon further inspection, 23 teams were removed from the dataset due to apriori criteria for team exclusion. For example, meetings planned for 30 min or less were deemed too brief for the TFAT to capture the expected observable behaviors and were subsequently omitted from the study. The remaining 110 MDHTs consisted of 35 mental health teams (32%), 33 medical teams (30%), 15 pediatric teams (14%), 10 surgical teams (9%), 8 oncology and palliative care teams (7%), 6 geriatric teams (5%), and 3 orthopedic teams (3%). Table 1 illustrates the demographic details of the teams.

Two independent raters observing the team meeting at its regularly scheduled time administered the TFAT. A pool of nine TFAT assessors working in rotating pairs conducted the assessments. In addition, all members of the MDHT present at the meeting were asked to complete a team experiences questionnaire assessing demographic characteristics and the four variables relating to job demands and job resources prior to leaving the meeting. Participants also were provided the option of completing the questionnaire outside of the meeting with an addressed reply-paid envelope. These participants were requested to complete and return the questionnaire within 24 hr. Response rates across teams ranged from 13% to 100% (mean response rate was 59%).

Measures

Team functioning.

Observational ratings of team functioning were measured using the 61-item version of the TFAT (Sutton et al., 2011). Each team behavior was rated on a 7-point behaviorally anchored rating scale, ranging from 1 (*Very poor—Performance on behavioral exemplar item clearly unacceptable with counterproductive behaviors that have negative outcomes or consequences*) to 7 (*Excellent—performance on behavioral exemplar item demonstrates mastery and could be used as a positive example or role model for others*). The TFAT consisted of 17 behavioral exemplar items relating to Clinical Planning (e.g., "Team shows an awareness of potential threats or changes in the environment"), 21 items relating to Executive Tasks (e.g., "Team is prepared with respect to materials and information"), and 23 items relating to Team Relations (e.g., "Team considers needs of others before acting"). Nine raters underwent a comprehensive 2-day frame-of-reference (FOR) training course in the use of the TFAT prior to data collection. FOR training has been shown to improve rating accuracy by reducing rater idiosyncrasies

Table 1. Descriptive Statistics of Team Characteristics ($N = 110$)

	Mean	*SD*	Range
Age (years)	38.03	5.33	26.67–57.00
Team size (people)	12.01	4.94	4.00–29.00
Team tenure (months)	32.31	22.56	1.50–144.00
Hospital tenure (months)	56.50	37.83	8.17–282.00
Occupational tenure (months)	152.67	56.84	34.67–322.00

(Bernardin, Buckley, Tyler, & Wiese, 2000; Uggerslev & Sulsky, 2008).

Organizational constraints.
Based on Peters and O'Connor's (1980) organizational constraints scale, team members were asked to rate how often they found it difficult to do their jobs on 11 items assessing organizational constraints, such as "Inadequate training." The rating scale was a 7-point scale, ranging from 1 (*never*) to 7 (*always*). Cronbach's α of .87 demonstrated good internal consistency. In line with Chan's (1998) recommendation for justifying data aggregation of individual responses to the group-level, within-group agreement using the r^*_{wg} statistic for multiple-item scales was calculated (Lindell & Brandt, 1999). Furthermore, the variance attributable to the group-level was examined using ICC (1), as well as the reliability of group member responses using ICC (2) (Bliese, 2000). The mean and median r^*_{wg} for the 110 groups were .83 and .87, respectively, indicating there was an acceptable level of within-group agreement (Lance, Butts, & Michels, 2006). ICC results showed acceptable ranges for organizational research (Klein et al., 2000) with ICC (1) and ICC (2) values of 0.03 and 0.18, respectively.

Time pressure.
Six items were adapted from Cousins et al.'s (2004) eight-item demands scale, with responses ranging from 1 (*never*) to 7 *(always*). Items were adapted to shift the referent of items to the team-level. An example item included "*Our team is pressured to work long hours*". Cronbach's α of .92 demonstrated a high level of internal consistency. There was a high level of within-group agreement (r^*_{wg} mean = .92 and r^*_{wg} median = .94). Furthermore, ICC (1) and ICC (2) values of .68 and .94, respectively, justified the aggregation of individual responses to the group level.

Leadership ambiguity.
Leadership ambiguity was assessed with three items developed for use in the present study and were modeled on the single-item measure developed by West et al. (2003). An example item included "There is no clear leader/coordinator in this team." Items were rated on a 7-point scale, ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The three items were internally consistent with a Cronbach's α of .92. Assessment of within-group agreement demonstrated high levels of agreement among group members (r^*_{wg} mean = .96 and r^*_{wg} median = .98). Furthermore, ICC (1) and ICC (2) values of 0.16 and 0.58, respectively, demonstrated acceptable levels of variance attributable to the group level.

Group potency.
Seven items were taken from Guzzo et al. (1993) eight-item measure, ranging from 1 (*to no extent*) to 7 (*to a very great extent*). An example item is "This team has confidence in itself." Cronbach's α of .96 demonstrated a high level of internal consistency of items in the scale. In support of aggregation to the group-level, there was high within-group agreement among group members (r^*_{wg} mean = .97 and r^*_{wg} median = .97). There also was high levels of variance attributable to the group, ICC (1) = 0.68, and group member responses were reliable, ICC (2) = 0.94.

Control variables.
In order to rule out alternative explanations for our findings, the following variables were controlled for: hospital size (either small or large [over 500 beds]), team size (number of team members present at the meeting), and team type. Team type was defined as the type of division in which the team was working (i.e., medical, surgical, pediatric, mental health, palliative care, oncology, acute rehabilitation).

Results

All statistical analyses conducted were performed using *MPlus 5.0*. Prior to analyses, individual-level responses were aggregated to the group level (i.e., averaging group member responses within each group) in order to statistically assess and describe relationships at the group level (Chan, 1998; Klein, Conn, Smith, & Sorra, 2001).

Initial Item Reduction and Assessment of Applicability

In order to reduce the number of items in the original TFAT, and to assess the applicability of items in the tool, three SMEs, who had extensive experience using the TFAT to rate MDHT functioning, were asked to assess whether items could discriminate between high- and low-functioning teams. In assessing the usability of items, SMEs also considered assessor feedback, including issues that were formally documented by assessors in a real-time TFAT feedback and instruction manual, as well as ongoing issues highlighted in assessor debriefing meetings. Discussions among SMEs led to recommendations to delete three items from Clinical Planning, three items from Executive Tasks, and six items from Team Relations for a variety of reasons. For example, "Team communicates and records time lines for planned actions in a centrally held document" was removed from the TFAT items, and instead, replaced with a series of tick boxes allowing the respondent to indicate what type of information is routinely collected and thereby providing more accurate and useful data.

Exploratory Factor Analyses

Based on the remaining set of 49 TFAT items, three EFAs using maximum likelihood estimation and oblimin rotation were performed with 14 items for the latent variable of Clinical Planning, 18 items for the latent variable of Executive Tasks, and 17 items for latent variable of Team Relations. Consistent with Tabachnick and Fidell's (2007) recommendations, the number of factor solutions was determined by eigenvalues for factor solutions above 1 in conjunction with scree-plot tests. The goodness of fit indices for the obtained factor solutions also were examined, along with investigating solutions that made theoretical sense.

Results of EFAs indicated that a three-factor solution for Clinical Planning had poor model fit, χ^2 (52, 110) = 147.01, $p < .001$, but acceptable fit indices, CFI = 0.91, Standardized Root Mean Square Residual (SRMR) = 0.04, Root Mean Square Error of Approximation (RMSEA) = 0.13, 90% CI = 0.11–0.15. Furthermore, a five-factor solution for Executive Tasks had good model fit, χ^2 (73, 110) = 91.65, $p = .069$, CFI = 0.96, SRMR = 0.02, RMSEA = 0.05, 90% CI = 0.00–0.08, while a three-factor solution for Team Relations had poor model fit, χ^2 (88, 110) = 139.62, $p < .001$, but acceptable fit indices, CFI = 0.97, SRMR = 0.03, RMSEA = 0.07, 90% CI = 0.05–0.10.

The factor structure for each of the three TFAT categories was determined based on item-factor loadings, such that item-loadings above 0.40 were considered as loading well onto the factor (Tabachnick & Fidell, 2007). Two items were deleted from Clinical Planning due to items cross-loading across subcategories. Five items were deleted from Team Relations, with three items cross-loading across subcategories, and two items lacking conceptual fit with other items in the subcategory factor. Based on the clustering of the remaining 42 items, SMEs determined the underlying constructs and provided descriptions for each of the 11 subcategory scales. Importantly, results of item-factor loadings also led to one change in the predicted factor structure, such that "Medical officer seeks information from team members" was changed from an item in the subcategory of "Case management" to an item in the subcategory of "Perception of elements in environment," which was conducted with SME consultations.

Single Congeneric Factor Models

Single congeneric factor models were performed as the first step (Joreskog, 1971). Tests are considered congeneric if, with the exception of measurement errors, the tests measure the same construct. Eleven separate single-congeneric factor models were conducted for each of the 11 TFAT subcategories, with each single-congeneric factor model depicting a single subcategory as the latent variable, and items as the manifest variables. Overall goodness of fit and model fit indices were inspected to evaluate the validity of the factor structures. However, because model fit indices are not provided for factor solutions with three or less items, the

squared multiple correlation (SMR) and standardized regression weights (β) are reported in these instances. If model fit was poor, it was improved by examining modification indices. Decisions regarding item deletions also were done in consultation with SMEs.

Model fit indices for the single-congeneric factor models ranged from acceptable to very good, with the exception of the subcategories of "Perception of elements in the environment" and "Case management." The five items in "Perception of elements in environment" demonstrated a poor model fit, χ^2 (5, 110) = 42.45, $p < .001$, CFI = 0.87, SRMR = 0.06, RMSEA = 0.18, 90% CI = 0.19–0.34. After examining modification indices and consulting with SMEs, item #4 ("Team acknowledges suboptimal outcomes") was deleted, and results with the remaining four items demonstrated good model fit, χ^2 (2, 110) = 9.14, $p = .010$, CFI = 0.97, SRMR = 0.03, RMSEA = 0.18, 90% CI = 0.07–0.31. The five items in "Case management" also demonstrated poor model fit, χ^2 (5, 110) = 41.14, $p < .001$, CFI = 0.89, SRMR = 0.06, RMSEA = 0.26, 90% CI = 0.19–0.33, but upon deletion of item #16 ("Team considers limitations /impact of outside services/caregivers"), model fit improved, χ^2 (2, 110) = 5.14, $p = .080$, CFI = 0.99, SRMR = 0.02, RMSEA = 0.12, 90% CI = 0.00–0.25. It also should be noted that the item "Identification and utilization of resources" yielded marginally acceptable fit indices. However, the SRMR and SMR statistics demonstrated that this subcategory is still an acceptable model. Table 2 illustrates the model fit statistics for the final 11 subcategory single-congeneric models. Across the 11 single-congeneric models, standardized path estimates of item loadings were all significant, ranging from 0.43 to 1.00. SMR coefficients also demonstrated that there was good average variance extracted from the items, ranging from 0.33 to 0.99. There was high internal consistency in the scale; Cronbach alphas ranged from .75 to .94.

Confirmatory Factor Analyses

Based on the factor structure established in the final 11 single-congeneric factor models, three separate CFA models for Clinical Planning, Executive Tasks, and Team Relations were conducted. In each of these CFA models, subcategories were depicted as latent variables,

Table 2. Model Fit Indices and Measurement Model Statistics for 11 Team Functioning Assessment Tool (TFAT) Subcategory Single Congeneric Factor Models

		Overall Model Fit			Model Fit Indices				Additional Statistics		
	No. of items	χ^2	*df*, *N*	*p*	CFI	SRMR	RMSEA	90% CI	SMR	β	α
1. Perception of elements in environment	4	9.14	2,110	.010	.97	.03	.18	.07–.31	.59–.78	.50–.74	.85
2. Case presentation	2	–	–	–	–	–	–	–	.66–.99	.81–1.00	.88
3. Case management	4	5.14	2,110	.080	.99	.02	.12	.00–.25	.48–.76	.70–.87	.86
4. Identification and utilization of resources	4	3.01	2,110	.001	.87	.06	.18	.19–.34	.66–.97	.43–.95	.87
5. Leadership decisiveness	4	4.23	2,110	.120	.99	.01	.10	00–.24	.55–.88	.74–.94	.93
6. Innovation climate	3	–	–	–	–	–	–	–	.48–.84	.69–.91	.87
7. Execution of meetings	3	–	–	–	–	–	–	–	.33–.95	.58–.97	75
8. Leadership inspiration	4	11.47	2,110	.003	.96	.03	.21	.10-.33	.63–.78	.79–.88	.89
9. Appropriate communication	3	–	–	–	–	–	–	–	.35–.74	.59–.86	.77
10. Appropriate use of authority and assertiveness	4	0.48	2,110	.786	1.00	.00	.00	.00–.12	.70–.85	.83–.92	.93
11. Consideration of others	4	4.90	2,110	.428	1.00	.01	.00	.00–.13	.65–.82	.91–.90	.94

with items as the manifest indicators. In order to confirm our predicted measurement model, items were allowed to freely load onto their respective subcategories, and covariances were estimated between subcategory latent variables.

Clinical planning.
Model fit indices demonstrated a good fitting model for the predicted factor structure of Clinical Planning, χ^2 (32, 110) = 57.74, $p = .004$, CFI = 0.96, SRMR = 0.05, RMSEA = 0.09, 90% CI = 0.05–0.12. All standardized regression estimates for the 10 items were significant, with items loading onto the predicted subcategories in Clinical Planning (see Figure 1).

Executive tasks.
The predicted measurement model of Executive Tasks yielded a good fitting model, χ^2 (125, 110) = 209.61, $p < .001$, CFI = 0.94, SRMR = 0.08, RMSEA = 0.08, 90% CI = 0.06–0.10. All standardized regression estimates for the 18 items were significant, with items loading onto the predicted subcategories in Executive Tasks (see Figure 2).

Team relations.
Model fit for Team Relations demonstrated a good fitting model, χ^2 (51, 110) = 89.52, $p = .001$, CFI = 0.97, SRMR = 0.05, RMSEA = 0.08, 90% CI = 0.05–0.11. All standardized regression estimates for the 12 items were significant, with items loading onto the predicted subcategories in Team Relations (see Figure 3).

Inter-rater Agreement and Reliability of TFAT Ratings

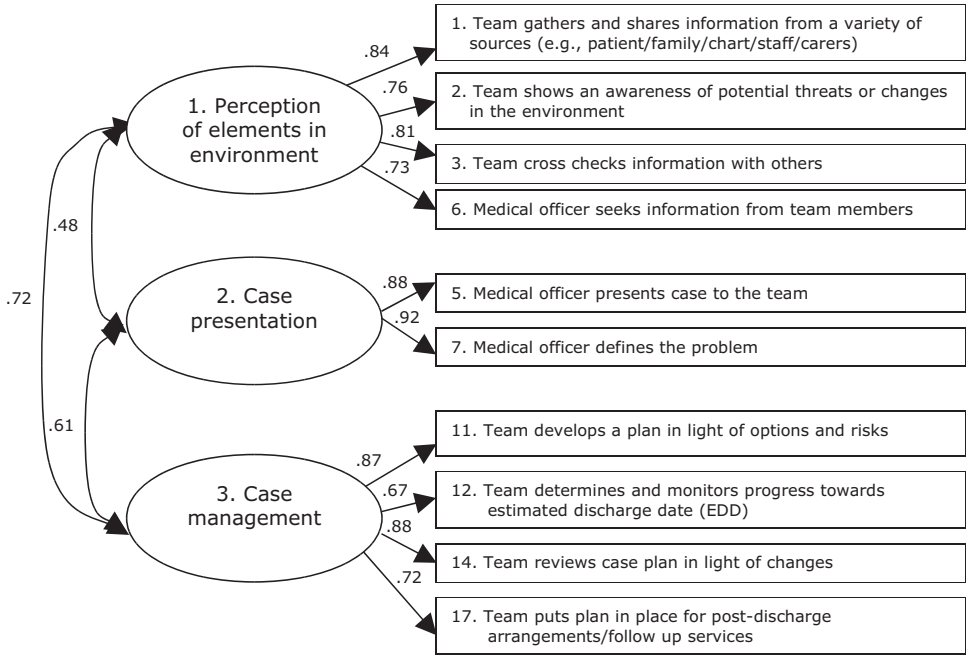
Consistent with Sutton et al. (2011), r^*_{wg} statistics were analyzed for each team in order to provide justification for aggregating the two raters' TFAT ratings, with r^*_{wg} values above .70, indicating good inter-rater agreement of TFAT scores (Lindell & Brandt, 1999). Interclass correlation coefficients of ICC (1) and ICC (2) also were used to determine the level of consistency and reliability of scores between raters for the purposes of aggregating the two assessors' TFAT ratings of teams. Table 3 describes the r^*_{wg}, ICC (1), and ICC (2) values for the 11 subcategory behavioral elements and higher-order categories in the TFAT. Given that the TFAT adopted a 7-point response scale format, the error variance was set to 4 when calculating r^*_{wg} values (see LeBreton & Senter, 2008). Mean r^*_{wgs} of two raters across 110 teams ranged from .72 to .96, indicating that there was moderate to strong agreement between raters on scores for each of the higher-order categories and behavioral elements (Lindell & Brandt, 1999). The ICC (1) values ranged from 0.68 to 0.89, and ICC (2) values were all above 0.70, indicating that there was high reliability between raters on TFAT ratings (Bliese, 2000).

Concurrent Validation

Our last demonstration of construct validation of the TFAT was conducted through concurrent validity testing, which involves correlating the TFAT variables of theoretical relevance obtained at the same time as TFAT ratings. Specifically, it was hypothesized that the TFAT observational ratings would be related to four job characteristic variables that were rated by teams as a whole: (1) organizational constraint, (2) time pressure, (3) leadership ambiguity, and (4) group potency. Descriptive information for the TFAT and job demands and job resources are shown in Table 4.

Hierarchical multiple regressions were conducted to assess the relationships between job demands and job resources, and Clinical Planning, Executive Tasks, and Team Relations (see Table 5). After controlling for the effects of hospital size, team size, and team type at Step 1, the job demand and job resource variables were entered in at Step 2 of the regression. Results demonstrated that over and above the effects of hospital size, team size, and team type, ratings of Clinical Planning were negatively associated with team experiences of organizational constraints and positively associated with group potency. The negative relationship between Clinical Planning and time pressure was approaching significance, and leadership ambiguity was unrelated to Clinical Planning. Executive Tasks was negatively related to organizational constraints and leadership ambiguity, and positively associated with group potency, although time pressure had no significant association with Executive Tasks. As predicted, Team Relations was negatively related to organizational constraints. The positive relationship between Team Relations and group potency also was approaching significance, whereas leadership ambiguity and time pressure had no significant relationship with Team Relations.

Figure 1. Confirmatory Factor Analysis (CFA) of Clinical Planning



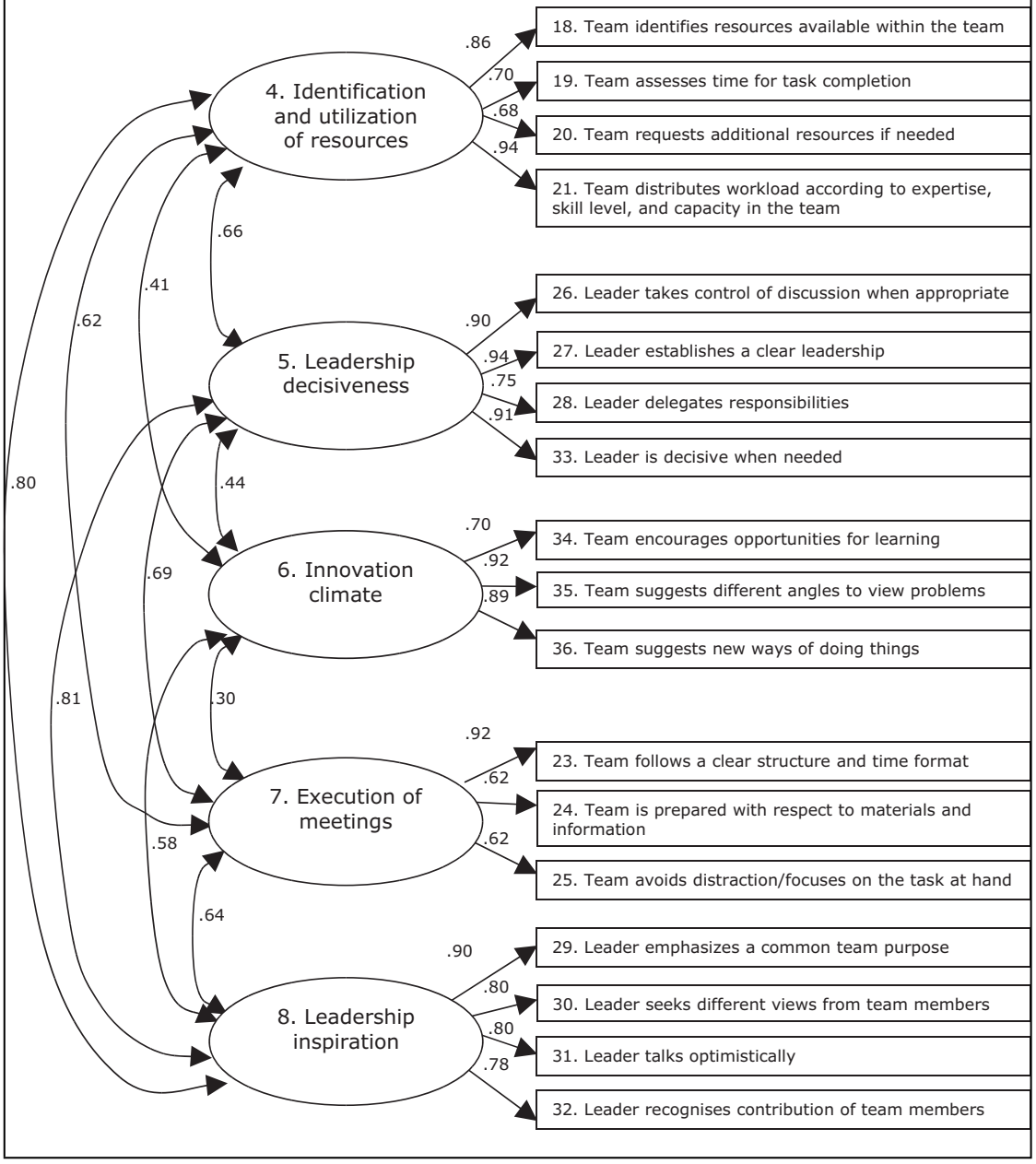
Influence of Hospital Size, Team Size, and Team Type

In order to assess the utility of the TFAT across a range of different hospital and team contexts, the extent to which TFAT ratings varied as a function of hospital size, team size, and team type were examined. First, independent sample *t*-tests demonstrated that hospital size had no significant effect on ratings of Clinical Planning, $t\,(108) = 0.04$, $p = .968$, Executive Tasks, $t\,(17.19) = -1.02$, $p = .321$, and Team Relations, $t\,(108) = 0.71$, $p = .477$. Second, regression analyses showed that team size was unrelated to ratings for Clinical Planning, $\beta = 0.02$, $p = .846$; Executive Tasks, $\beta = -0.03$, $p = .797$; and Team Relations, $\beta = -0.09$, $p = .347$. Third, a series of one-way ANOVAs demonstrated that team type had no significant effect on ratings of Executive Tasks, $F\,(6, 103) = 2.04$, $p = .072$, and Team Relations, $F\,(6, 103) = 1.60$, $p = .154$, although team type had an effect on ratings of Clinical Planning, $F\,(6, 103) = 2.36$, $p = .036$. Specifically, Bonferroni-adjusted *t*-tests showed that geriatric teams ($M = 5.87$; $SD = 0.33$) scored higher than medical teams ($M = 4.90$; $SD = 0.60$) on Clinical Planning, $t'\,(103) = 3.17$, $p = .043$. These results illustrate that with the exception of differences between geriatric and medical teams on Clinical Planning, TFAT ratings did not differ as a function of hospital size, team size, or team type. Overall, there is support for the utility of the tool across a range of team and hospital contexts.

Discussion

The primary goal of this paper was to validate and present a shortened version of the TFAT that can identify, measure, and evaluate nontechnical skills relating to team functioning in a "non-crisis" ward-based MDHT setting (Sutton et al., 2011). Based on content validity testing with SMEs, considerations of usability and applicability of each item, as well as results from testing the factor structure, the length of the TFAT was reduced from 61 to 40 behavioral items. Importantly, from this revision, the TFAT was adapted to include a series of tick boxes that aimed to capture (1) the mechanism (if any) that teams used to record meeting outcomes, (2) the physical accessibility of this outcome data, and (3) the information that is routinely collected in the document or tool. Providing support for the factor structure of the revised TFAT, our findings showed that the 40 items loaded onto their respective 11 subcategories, and the 11 subcategories loaded onto the predicted

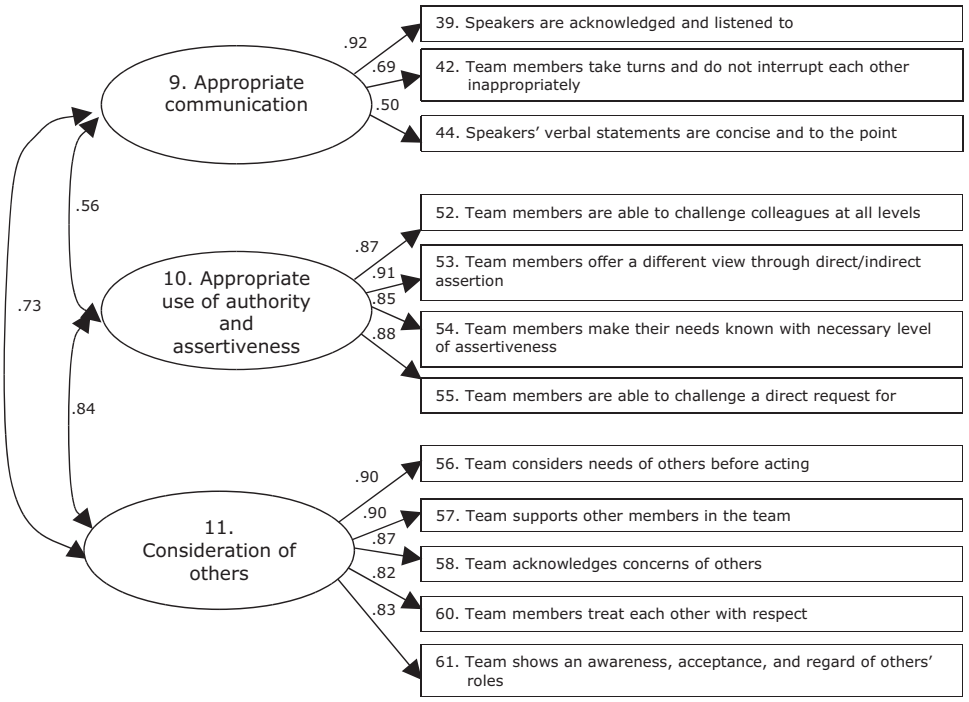
Figure 2. Confirmatory Factor Analysis (CFA) of Executive Tasks



three core categories of Clinical Planning, Executive Tasks, and Team Relations.

In relation to construct validation, Hypothesis 1 was fully supported in which high levels of organizational constraints were associated with low observer ratings for all three TFAT categories. Hypothesis 2 was partially supported, such that teams experiencing high levels of time pressure were rated poorly on Clinical Planning. However, there was no relationship between time pressure and Executive Tasks or Team Relations which could be attributable to the fact that insufficient time for workload demands usually stem from demands outside of the meeting in the delivery of patient care. Thus, time pressure may have a more direct effect on team functioning behaviors relating to how teams plan and coordinate the delivery of patient care (Clinical Planning), as opposed to behaviors that are specific to during the team meeting, such as how the meeting is conducted (Executive Tasks) or how team members

Figure 3. Confirmatory Factor Analysis (CFA) of Team Relations



interact with each other in the meeting (Team Relations).

Similarly, Hypothesis 3 also was partially supported. High levels of leadership ambiguity were associated with low Executive Tasks ratings, although there was no relationship between leadership ambiguity and Clinical Planning or Team Relations. The finding that leadership ambiguity was only negatively related to Executive Tasks (and not Clinical Planning or Team Relations) may be because Executive Tasks includes subcategories of "Leadership decisiveness" and "Leadership inspiration," which contain items that describe effective behaviors of the leader. It may be that leadership ambiguity is not related to how well teams engage in clinical planning behaviors, or how well team members relate to each other because these categories are not conceptually related to a lack of leadership in the team.

Finally, Hypothesis 4 was fully supported, with high levels of group potency being associated with high ratings for all three categories in the TFAT. Overall, our findings are consistent with the JD-R model (Bakker & Demerouti, 2007; Demerouti et al., 2001), proposing that job demands stemming from the organization, the task, and the team can reduce capacity for engaging in effective team functioning behaviors, whereas job resources can motivate people to engage in effective team functioning behaviors. Taken together, our findings provide preliminary support for the construct validity of the refined TFAT's factor structures, and the applicability of the TFAT as a domain-specific behavioral marker system for evaluating the nontechnical skills of "non-crisis" MDHTs.

Our study also demonstrated the reliability of items within each of the 11 subcategories, and moreover, it demonstrated that independent raters can be trained to produce reliable ratings on the revised TFAT. However, the use of FOR training is recommended to reduce the impact of rater biases and error on TFAT ratings (Bernardin et al., 2000; Uggerslev & Sulsky, 2008). One of the future challenges is how to make this type of training available to a wide audience in conjunction with the TFAT tool. It could be argued that providing and interpreting feedback to teams on their uptake of nontechnical skills over time will inevitably expose team members to an implicit form of FOR training, focusing on identification and discussion of the behaviors required for effective ward-based MDHTs. However, it is recommended that training should be employed to expose

Table 3. Inter-rater Agreement (r^*_{wg}) and Inter-rater Reliability (ICC 1 and ICC 2) of Team Functioning Assessment Tool (TFAT) Ratings across 110 Teams

	r^*_{wg}		ICC	
	Mean	**Median**	**ICC(1)**	**ICC(2)**
Clinical planning	0.85	0.89	0.83	0.91
1. Perception of elements in environment	0.91	0.94	0.69	0.82
2. Case presentation	0.92	0.97	0.82	0.90
3. Case management	0.89	0.94	0.84	0.92
Executive tasks	0.72	0.73	0.89	0.94
4. Identification and utilization of resources	0.90	0.94	0.78	0.88
5. Leadership decisiveness	0.91	0.94	0.83	0.91
6. Innovation climate	0.86	0.93	0.86	0.93
7. Execution of meetings	0.90	0.94	0.73	0.84
8. Leadership inspiration	0.90	0.94	0.80	0.89
Team relations	0.91	0.92	0.77	0.87
9. Appropriate communication	0.92	0.94	0.68	0.81
10. Appropriate use of authority and assertiveness	0.96	0.98	0.69	0.82
11. Consideration of others	0.93	0.95	0.73	0.85
Overall team functioning (TFAT)	0.79	0.81	0.87	0.93

Table 4. Descriptive Statistics and Bivariate Correlations between Team Functioning Assessment Tool (TFAT) and Job Demands and Resources

	M	*SD*	1	2	3	4	5	6
1. Clinical planning	5.10	0.71						
2. Executive tasks	4.86	0.73	0.84**					
3. Team relations	5.26	0.65	0.78**	0.72**				
4. Organizational constraints	2.99	0.52	−0.14	−0.16	−0.18†			
5. Time pressure	3.83	0.64	−0.14	−0.06	−0.13	0.40***		
6. Leadership ambiguity	2.43	0.77	−0.09	−0.24*	−0.01	0.20*	0.03	
7. Group potency	5.13	0.57	0.14	0.14	0.07	−0.36***	−0.14	−0.40***

***$p < .001$; **$p < .01$; *$p < .05$; †$p < .10$.

raters to an explicit discussion of, and progress toward, a common FOR.

The TFAT also was tested across hospital size, team size, and team type. Ratings generally did not differ as a function of hospital and team characteristics, with the exception that geriatric teams scored better on Clinical Planning than medical teams. Although some clinicians may be quick to explain this finding through a greater opportunity for geriatric teams to demonstrate nontechnical skills (because geriatric teams tend to engage in longer case conferencing than medical teams), it should be noted that teams scheduled to meet for 30 min or less, and 3 hr or more, were excluded from our analyses. Thus, higher scores on Clinical Planning for geriatric teams has more to do with a collaborative approach to Clinical Planning and individual preparation for information sharing than with ALOS or duration of team meetings. Overall, the revised TFAT can be applied across a range of MDHT contexts to assess and evaluate team functioning effectiveness.

Limitations and Future Research Directions

In relation to our concurrent validity testing, measures for job demands and job resources variables were adapted from validated measures. Although the full set of items were not employed in some of the scales due to questionnaire length restrictions (i.e., group potency and time pressure), the highest factor-loading items were included in the selection of items. Furthermore, while aggregation statistics

Table 5. Relationships among Team Functioning Assessment Tool (TFAT) Scores and Job Demands and Resources

	TFAT Scores							
	Clinical Planning		Executive Tasks		Team Relations		Overall	
Job demands and resources	β	*p*	β	*p*	β	*p*	β	*p*
Organizational constraints	−0.22	0.036*	−0.24	0.027*	−0.27	0.013*	−0.26	0.014*
Time pressure	−0.18	0.084†	−0.10	0.311	−0.14	0.171	−0.14	0.167
Leadership ambiguity	−0.10	0.331	−0.23	0.017*	−0.03	0.744	−0.16	0.109
Group potency	0.29	0.014*	0.27	0.020*	0.20	0.091†	0.27	0.019*

Note. Team size, team type, and hospital size controlled at Step 1.

justified aggregation of the job characteristic variables to the group-level, it should be noted that ICC statistics for organizational constraints was the lowest but was, nevertheless, comparable to other aggregation statistics seen in organizational research (see Bliese, 2000). Organizational constraints in the healthcare context may exist beyond the immediate team experience, with team members having different experiences of organizational constraints, depending on their professional roles in the hospital. Furthermore, it was beyond the scope of this study to collect patient outcome data, and future research should aim to demonstrate the predictive validity of the TFAT on patient outcomes. Although the present paper established the factor structure for each of the three core categories, it should be noted that it did not demonstrate that the TFAT as a whole comprises three empirically distinct factors and future research should aim to replicate these findings. Although the sample size was adequate for the statistical analyses undertaken in this study, MacCallum and Austin (2000) recommend employing over 220 cases for conducting CFAs with 11 manifest indicators. Recognizing this limitation, it is suggested that future researchers assess the factor structure of the TFAT with a larger sample of teams.

In line with research on behavioral marker tools (Klampfer et al., 2001), the TFAT was designed to assess a specific team context; that is, ward-based MDHTs. The experience gained in developing the TFAT and applying CRM to healthcare teams beyond the crisis environment can serve to inform the development process for behavioral marker systems for use with other types of clinical or administrative teams within healthcare. However, such behavioral marker tools will, by necessity, reflect a unique set of technical and nontechnical skills for effective team behavior. Thus, future research should adapt the TFAT to suit clinical contexts beyond ward-based teams, such as extended rehabilitation and community based care, as well as teams without direct clinical contact working in healthcare. Furthermore, careful consideration should be given to the release and application of TFAT data. Data obtained from TFAT observations also should be interpreted within a program of planned change and support for participating teams.

Widespread interest exists in the area of team effectiveness within the modern healthcare environment, with multiple initiatives aimed at improving access to services and quality of care ("Clinical Services Redesign Program", 2011). One of the future challenges is to provide adequate links between existing training opportunities, redesign initiatives, and research projects with the goal to ensure that potential gains are realized and spread throughout large healthcare organizations. The TFAT contributes to all three areas of work. Specifically, assessment of performance levels in and across MDHTs using the TFAT will identify areas of strengths and areas of improvements regarding the functioning of teams, and allow for training interventions to be tailored to the needs of the team. Subsequently, this assessment will permit meaningful comparisons between baseline preintervention measures and postintervention measures of redesign initiatives relating to healthcare delivery. Finally, with respect to developing evidence-based training interventions, the TFAT provides a usable, reliable, and valid measure for researchers to empirically assess whether the planned interventions improve skills and behaviors for effective team functioning.

Acknowledgment

This study was supported by an Australian Research Council Grant (LP0882135) awarded to the first, third, and fourth authors.

References

Bakker, A. B., & Demerouti, E. (2007). The job demands-resources model: State of the art. *Journal of Managerial Psychology, 22*, 309–328.

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A reconsideration of strategies in rater training. *Research in Personnel and Human Resources Management, 18*, 221–274.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Carter, S., Garside, P., & Black, A. (2003). Multidisciplinary team working, clinical networks, and chambers; opportuntities to work differently in the NHS. *Quality & Safety in Health Care, 12*, i25–i28.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*, 234–246.

Clinical Services Redesign Program Home. (2011). In *Centre for Heathcare Improvement*. Retrieved November 21, 2011, from www.health.qld.gov.au/chi/CSRP/default.asp.

Cousins, R., MacKay, C. J., Clarke, S. D., Kelly, C., Kelly, P. J. & McCaig, R. H. (2004). Management standards and work-related stress in the UK: Practical development. *Work Stress, 18*(2), 113–136.

Demerouti, E., Bakker A.B., Nachreiner, F., Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology, 86*, 499–512.

Firth-Cozens, J. (2001). Cultures for improving patient safety through learning: The role of teamwork. *Quality in Healthcare, 10*, ii26–ii31.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' non-technical skills (ANTS): Evaluation of a behavioral marker system. *British Journal of Anaesthesia, 90*, 580–588.

Flin, R., & Maran, N. (2004). Identifying and training non-technical skills for teams in acute medicine. *Quality & Safety in Health Care, 13* (Suppl 1), i80–i84.

Guzzo, R. A., Yost, P. R., Campbell, R. J., & Shea, G. P. (1993). Potency in groups: Articulating a construct. *British Journal of Social Psychology, 32*, 87–106.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–132.

Kelly, J. R. & Loving, T. J. (2004). Time pressure and group performance: Exploring underlying processes in the attentional focus model. *Journal of Experimental Social Psychology, 40*, 185–198.

Klampfer, B., Flin, R., Helmreich, R. L., Hausler, R., Sexton, B., Fletcher, G., et al. (2001). Enhancing performance in high risk environments: Recommendations for use of behavioral markers. *Daimler-Benz Shiftung: GHIRE*, Retrieved November 20, 2007, from www.abdn.ac.uk/iprc/papers%20reports/Ants/GIHRE21_rec_for_use_of_beh_markers.pdf.

Klein, K. J., Bliese, P. D., Kozolowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., et al. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology, 86*, 3–16.

Kosnik, L. K. (2002). The new paradigm of crew resource management: Just what is needed to reengage the stalled collaborative movement? *Joint Commission Journal for Quality Improvement, 28*, 235–241.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*, 202–220.

LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815–852.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance if a test: A comparison of the CVI, T, rWG(J), and r*WG indexes. *Journal of Applied Psychology, 84*, 640–647.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modelling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Miller, L. A. (2005). Patient safety and teamwork in perinatal care: resources for clinicians. *Journal of Perinatal & Neonatal Nursing, 19*, 46–51.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. *Academy of Management Review, 5*, 391–397.

Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job satisfaction and strain: Interpersonal Conflict at Work Scale, Organisational Constraints Scale, Quantitative Workload Inventory, and Physical Symptoms Inventory. *Journal of Occupational Health Psychology, 3*, 365–367.

Sutton, G., Liao, J., Jimmieson, N. L., & Restubog, S. L. D. (2011). Measuring multidisciplinary team effectiveness in a ward-based healthcare setting: Development of the team functioning assessment tool. *Journal of Healthcare Quality, 33*, 10–24.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*, 5th ed. Boston, MA: Allyn and Bacon.

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*, 711–719.

West, M. A., Borill, C. S., Dawson, J. F., Brodbeck, F., Shapiro, D. A., & Howard, B. (2003). Leadership clarity and team innovation in healthcare. *The Leadership Quarterly, 14*, 393–410.

Wilson, K. A., Burke, C. S., Priest, H. A., & Salas, E. (2005). Promoting health care safety through training high reliability teams. *Quality and Safety in Health Care, 14*, 303–309.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2004). Critical thinking: Non-technical skills in surgery. *Surgeons' News, 3*, 75–76.

Authors' Biographies

Gigi Sutton, PhD (OrgPsych), MBA, BOccThy (Hons), is the Manager for Team Effectiveness within the Clinical Skills Development Service of Queensland Health, Australia. She is interested in improving the quality of working

life for employees within large multidisciplinary teams, and associated organizational and patient-related outcomes.

Jenny Liao is a PhD student (OrgPsych), studying in the School of Psychology at The University of Queensland. Her research interest is concerned with how teams effectively coordinate expertise in the workplace, focusing on the development of transactive memory systems, and its effects on team-related outcomes.

Nerina L. Jimmieson is an Associate Professor in the School of Psychology at The University of Queensland (Australia). Employing a range of methodologies, her research interests are concerned with stress and coping in the workplace and employee adjustment to organizational change. Her research is well-supported by national competitive grants and industry partners, with publications appearing in journals and books in the areas of management and organizational psychology.

Simon Lloyd D. Restubog, PhD (OrgPsych), is Deputy Director for Research and Professor of Management in the Research School of Management at The Australian National University. His research interests include antecedents and maintenance of psychological contracts, measurement and prediction of destructive leadership and workplace deviance, and the role of support and barriers to career development in work and non-work contexts.

For more information on this article, contact Gigi Sutton at gigi_sutton@health.qld.gov.au or Jenny Liao at jennylia@psy.uq.edu.au.