The Joint Commission Journal on Quality and Patient Safety 2024; 50:95–103

The TeamSTEPPS for Improving Diagnosis Team Assessment Tool: Scale Development and Psychometric Evaluation

Kisha J. Ali, PhD, MS; Christine A. Goeschel, ScD, MPA, RN; Melissa M. Eckroade, MHA; Katie N. Carlin, MBA; Monika Haugstetter, MHA, MSN, RN, CPHQ; Margie Shofer, BSN, MBA; Michael A. Rosen, PhD, MA

Introduction: One in three patients is affected by diagnosis-related communication failures. Only a few valid and reliable instruments that measure teamwork and communication exist, and none of those focus on improving diagnosis. The authors developed, refined, and psychometrically evaluated the TeamSTEPPS® for Improving Diagnosis Team Assessment Tool (TAT), which assesses diagnostic teamwork and communication in five critical teamwork domains and can be used to identify strengths and opportunities for improvement and monitor performance.

Methods: The TAT was administered as a cross-sectional survey to 360 health professionals across nine diverse US health systems. Content and construct validity were evaluated through pilot implementation and subject matter expert review. Reliability and internal consistency were assessed with Cronbach's alpha. To understand sources of variation in TAT scores and assess the tool's consistency across diverse health care organizations, generalizability theory (G-theory) was used. Best practices in screening for careless responding identified participants with random or nonvarying responses.

Results: Analyses indicated strong support for the tool. Content validity findings indicated that the TAT encompassed relevant diagnostic improvement teamwork and communication content. Construct validity, evaluated through pilot implementations, demonstrated the tool's effectiveness in assessing teamwork categories. Reliability analyses confirmed the TAT's internal consistency, with an overall Cronbach's alpha of 0.97. Each dimension of the TAT exhibited good reliability coefficients, ranging from 0.83 to 0.95. G-theory analysis showed that variations in TAT scores were primarily attributed to respondents (28.0%) and scale dimensions (59.6%); both are desirable facets of variation. Further, examination of careless respondents ensured the accuracy and quality of the results, enhancing the TAT's credibility as a valuable diagnostic improvement tool.

Conclusion: Psychometric evaluation demonstrated that the TAT is a reliable and valid instrument for assessing teamwork and communication among and across diagnostic teams. The TAT adds a novel, evidence-based, psychometrically sound measurement tool to help advance diagnostic teamwork and communication to improve patient care and outcomes.

Diagnostic error is the failure to establish an accurate and timely explanation of the patient's health problem or to communicate that explanation to the patient, leading to delayed, wrong, or missed diagnosis.[1] There is urgent interest in the measurement and prevention of diagnostic errors, as evidence shows that one in three patients experiences a diagnostic error firsthand and that diagnostic-related communication failures occur across all care settings (that is, outpatient, inpatient, and emergency department).[1–3] Diagnosis is a collaborative effort.[2] Patients are safer and receive higher-quality care when providers work as an effective team.[4] According to The Joint Commission, one of the most reviewed sentinel events is delay in treatment, inclusive of communication failures and misdiagnosis.[5] Breakdowns in communication and teamwork are system failures that potentially lead to patient harm, and consequently teamwork interventions are one identified strategy for improving diagnostic processes.

It is imperative to have evaluation instruments that are reliable and valid to enable peer-comparator and consistent historical self-comparator data as well as accurate benchmarking and setting targets. However, few valid and reliable teamwork measures exist in health care.[6] Teamwork interventions use valid assessment tools to help organizations understand improvement opportunities and progress over time; however, no such tools that specifically target teamwork in service of improving diagnostic processes currently exist.

In this study, we address the need for a valid and reliable teamwork and communication measurement instrument to help improve patient diagnosis with the creation and psychometric evaluation of the TeamSTEPPS® for Improving Diagnosis Team Assessment Tool (TAT). The TAT is adapted from the Team Performance Observation Tool (part of TeamSTEPPS 2.0, developed by the Agency for

1553-7250/$-see front matter

https://doi.org/10.1016/j.jcjq.2023.08.009

Healthcare Research and Quality [AHRQ] and the US Department of Defense).[7] All indicators of the Team Observation Tool were modified to focus on diagnostic teamwork, creating the new indicators of the TAT, which then underwent content validity examination. Using data collected across seven diverse health care organizations for the TAT, we report internal reliability, parallel forms reliability, and results of a generalizability study (G-study) assessing individual respondents, TAT scale dimensions, and sites of data collection as important facets of variation.[8] With psychometric evaluation, the TAT becomes a measurement instrument to assess changes over time and introduce options for program evaluation and comparability in diagnostic teamwork and communication.

METHODS

Study Design

A parallel instrument survey to improve teamwork and communication related to patient diagnosis was administered cross-sectionally, using simple random sampling, to diverse health systems and various members of patient care teams, with and without quality improvement knowledge, across the United States.

Ethical Considerations

The TeamSTEPPS for Improving Diagnosis Team Assessment Tool and supplemental organizational demographic survey were approved by the US Office of Management and Budget (OMB No. 0935-0262, Exp. Date 06/30/2024) for implementation across nine health systems, and judged exempt by the MedStar Health Institutional Review Board (ID: MOD00011147, Review Date: 07/14/2022).

Scale Development

Under a Diagnostic Safety Capacity Building contract, AHRQ contracted the MedStar Institute for Quality and Safety (MIQS) to create resources for improving patient diagnosis and to psychometrically evaluate the TAT. The TAT was designed as part of the TeamSTEPPS for Diagnosis Improvement Course (a toolkit resource developed as part of the larger AHRQ contract). The TAT assesses the maturity phase of diagnostic teams in five critical teamwork categories: Team Structure, Communication, Leadership, Situation Monitoring, and Mutual Support.[9] It identifies strengths and opportunities for improvement in communication, directing teamwork priorities, developing action plans, and monitoring improvement.[9] The TAT was designed as a unit-level assessment, completed individually by members of the diagnostic team, with the goal of helping them reflect on their current teamwork and communication practices. When the maturity phase is assessed with an aggregate score, diagnostic teams can guide their improvement efforts using the TeamSTEPPS for Diagnosis Improvement Course, which has content linking to each domain on the TAT.

Prior to psychometric evaluation, precursor psychometric testing was conducted on the TAT (January to September 2021).[10–12] First, the wording of items was reviewed to ensure the appropriate Flesch-Kincaid Readability Statistics. Second, subject matter experts in teamwork and diagnostic safety ($N = 8$ reviewers) reviewed the TAT item content to ensure that all relevant (and no extraneous) diagnostic improvement teamwork and communication program content was included. Third, a subject matter expert panel composed of national content experts ($N = 7$ reviewers), outside the project team responsible for its development, provided formative item feedback as an additional assessment of content validity of TAT items and dimensions. Fourth, feedback was solicited from end users in practices ($N = 41$ users) on the TAT. Users did not complete the scale, but commented on the TAT content. Fifth, the TAT was further refined through pilot implementation at sites implementing the TeamSTEPPS for Diagnosis Improvement Course ($N = 41$ users, $N = 14$ diverse teams across the United States).

Subsequently, an alternate test instrument, also known as a parallel form, was developed (November 2021 to March 2022), pulling ideas from existing safety culture teamwork tools, to develop parallel indicators for the items in the TAT.[13–16] A parallel item was developed and mapped to each item of the original TAT (see Appendix 1, available in online article). Use of the parallel form allows for a rigorous assessment of the original TAT's structure (in other words, high correlations between corresponding domains of the equivalent scales are evidence of validity).

Recruitment

Figure 1 depicts the site recruitment and selection process. Recruitment of participating health systems was conducted nationwide (United States) in May 2022 via an AHRQ listserv announcement, a Society to Improve Diagnosis in Medicine listserv posting, MedStar Health listserv e-mails, and solicitation via social media platforms. Interested health systems ($N = 45$) sent a response e-mail to an MIQS point of contact. Eligibility was determined by health system size, type of organization (for example, private, size, rural, academic), and geographic region to ensure that an equitable sample of health care organizations were selected to participate. A panel of experts ($N = 5$ reviewers) involved in the development of the instrument selected participating organizations ($N = 9$ health systems) via a consensus process. Table 1 provides organizational demographic information for the participating facilities.

Study Population, Including Eligibility and Exclusion Criteria

Each health system received a single, unique, electronic survey link to the TeamSTEPPS for Improving

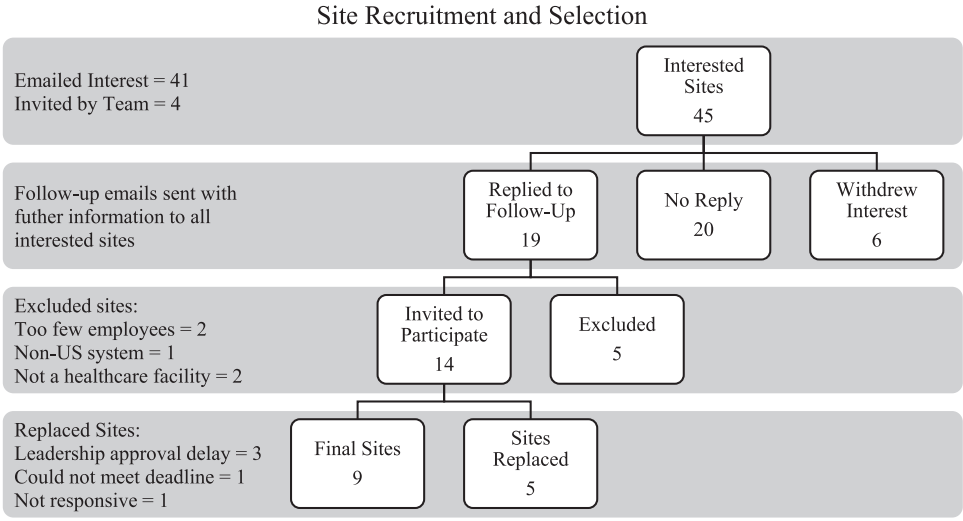

Figure 1: This figure depicts the site recruitment and selection process.

Table 1. Site Characteristics of the Final Nine Sites

Site Demographics	Total Sample ($N = 9$)	
	n	Frequency (%)
US Census Region		
Midwest	3	33.3
Northeast	1	
Southeast	3	33.3
Southwest	0	
West	1	11.1
Multi-Region	1	
Community Type		
Rural	2	
Suburban	3	33.3
Urban	1	
Multi-Community	3	33.3
Part of a Health Care System		
Yes	7	77.8
No	1	
N/A	1	11.1
Number of Beds		
> 50	2	22.2
50–500	1	
501–1,000	3	33.3
> 1,000	2	
N/A	1	11.1
Facility or Organization Type		
Health Care System	2	22.2
Acute Care Hospital	3	
Academic or Teaching Hospital	3	33.3
Alumni Association	1	

N/A, not applicable.

Diagnosis Team Assessment Tool to distribute within their organization. Individual responses within each organization were anonymized. The estimated completion time was 10 minutes. To tackle the prevalent and broadscale problem of preventable diagnostic errors, the TeamSTEPPS for Diagnosis Improvement Course[9]—which includes this TAT as part of its toolkit[17]—has recently (2021) shifted the mindset and culture of diagnostic care from only the clinical diagnostic staff to any staff who directly interacts with a patient, as errors affecting missed, delayed, inaccurate, and timely diagnosis can be attributed to any point in the care process from admission to discharge (for example, inpatient registration, administration, nurse, physician, ancillary care, lab, radiology, pharmacy). This new diagnostic patient safety culture mindset[18–20] shifts teamwork[21–24] in the diagnostic process to collaboration of a clinician with the patient, the patient's family members, and many other health care professionals; essentially, everyone who touches the patient plays a role on the diagnostic team, as facilitating teamwork among these individuals is critical to avoid failures in the health care processes that lead to preventable diagnostic errors.[9,18–20,25,26] In congruence with this new diagnostic team definition, a broad range of staff who had direct patient interaction was invited to complete the survey (for example, clinicians, trainees, technicians, registration staff, nurse aides, pharmacists, respiratory therapy).[9,17] Having a background in TeamSTEPPS or health care quality was not required. Instructional information provided to organizations emphasized that the results of the survey were not being collected or shared, but rather analysis was being performed to determine psychometric correlation on indicators within the instrument.

All care settings (ambulatory, surgical, inpatient) providing direct patient care from the participating health systems were invited to participate; as this is a culture-based teamwork and communication assessment, it is health care setting agnostic. Hospitals, health systems, and other health care settings eligible to participate designated an organizational point of contact and were sent paperwork for a stipend payment, the organizational demographic survey for one-time completion, and a unique organizationwide

link to complete the survey. Status updates of completion rates were sent during administration. The assessment period was July to August 2022. The target for each health care setting was to complete 100 surveys. A $1,000 stipend was provided to the health system as a thank you for participation after 60 completed surveys. The recommendation for psychometric evaluation using parallel forms is at least 10 participants for each scale item, with simulation study using different sample sizes illustrating a minimum sample of 300 surveys required to observe acceptable comparability of patterns, and replication being required if the sample size is < 300.[27] Our psychometric evaluation sample target was $N = 350$ completed surveys, which is well over the $N = 300$ surveys needed for parallel forms psychometric evaluation, to firmly ensure external validity and the robust statistical power of the data analysis.[7,11,12,27–33]

Statistical Analysis

A careless respondent analysis was conducted on the combined original and parallel forms of the TAT. See Appendix 2 for details.[34–37] To assess internal consistency, we calculated Cronbach's alpha for all TAT items collectively, as well as for items of each domain separately. Cronbach's alpha ranges from zero to one, with larger numbers indicating higher levels of reliability.[38] Although strict cutoffs for interpreting coefficient alpha is not recommended, the heuristic of a coefficient alpha ≥ 0.8 is generally accepted as sufficiently reliable for applied research purposes.[29] To test parallel forms reliability,[39] Pearson correlation coefficients were calculated for the matching domain and full scale scores from the original and parallel tests. Larger correlation coefficients are interpreted as higher degrees of reliability. Generalizability theory (G-theory) was used to understand sources of variation in TAT scores. G-theory is a rigorous psychometric approach employing analysis of variance (ANOVA) methods for quantifying systematic variance from multiple sources at once.[40–42] G-theory is applied here to evaluate the degree to which the TAT domains are unique and the degree to which the TAT may be useful in identifying unique teamwork needs across organizations. All analyses were conducted in R v. 4.1.3[43] using the careless[37] and G-theory packages.[44]

RESULTS

Table 1 illustrates that the psychometric evaluation of the nine final health care organizations represented a diverse distribution across various demographic factors. In terms of US Census Regions, the Midwest and Southeast were represented by three sites (33.3%) each; the Northeast, West, and multi-region each had one site (11.1% each). Community types spanned rural (22.2%), suburban (33.3%), urban (11.1%), and multi-community (33.3%). The majority of sites (77.8%) were part of a health care system, with a smaller proportion not affiliated (11.1%) or listed as not available (11.1%). Regarding the number of beds, the distribution ranges from less than 50 (22.2%) to greater than 1,000 (22.2%), with various sites falling into the 50 to 500 (11.1%), 501 to 1,000 (33.3%), or not available (11.1%) categories. Finally, facility or organization types included health care systems (22.2%), acute care hospitals (33.3%), academic or teaching hospitals (33.3%), and an alumni association (11.1%).

Of the nine participating health care organizations, there was attrition of one organization, and completion of only one survey by another organization; these sites were excluded from analysis. Of the remaining seven organizations, there were 758 incomplete survey responses, which were also excluded from analysis. Ultimately, 360 fully completed surveys were used for psychometric evaluation. Table 2 provides descriptive statistics for data by TAT dimensions and items. Reliability and generalizability analyses were similar when conducted with the full sample ($N = 360$) and a careless respondent screening sample ($n = 206$). To add another dimension of methodological rigor and precision in findings, analyses were conducted with the 206 responses remaining after the careless respondent screening. See Appendix 2 for details of careless respondent analyses and Appendix 3 for results of the following analyses conducted with the full 360 respondents.

Internal and Parallel Forms Reliability

As shown in Table 2, the overall reliability was high for the full TAT (Cronbach's alpha = 0.97), as were reliabilities for each dimension, ranging from 0.83 for Team Structure to 0.95 for the Leadership domain.[38,39] Parallel forms reliability was very high (r = 0.97) for the overall TAT scale, as well as for each dimension, ranging from 0.86 for Team Structure to 0.91 for Leadership (Table 2).

Generalizability of Study Findings

Table 3 details results of the G-study and presents the percentage of variance in scores associated with the tested facets. The individual respondent was the object of measurement and accounted for 28.0% of variance in scores. The TAT dimension accounted for an additional 59.6% of score variance, while the site (1.2%) and a site by dimension interaction (0.2%) accounted for substantially less variance. The remaining 11.0% of variance was unaccounted for (that is, either random error or associated with unaccounted-for facets).

DISCUSSION

With the TAT, participants complete self-assessment ratings to collectively identify strengths and opportunities for improving in unit-based teamwork, setting priorities, and

Table 2. Internal and Parallel Forms Reliability Estimates for All Items and by Subdimension

Dimension	Number of Items*	Mean (SD)	Cronbach's Alpha†	Parallel Forms Reliability
Overall TAT Scale	**25**		**0.97**	**0.97**
Team Structure	*5*	*3.8 (0.78)*	*0.83*	*0.86*
Each team member can identify all members of a diagnostic team (for example, patients, families, providers, radiology and lab personnel, other staff, support services).	–	3.6 (1.04)	0.80	–
All team members recognize the roles and responsibilities of each member of the diagnostic team.	–	3.7 (0.96)	0.78	–
Team members use defined communication tools (for example, SBAR, call-outs, check-backs, handoff techniques) to facilitate the diagnostic process.	–	3.6 (1.06)	0.78	–
Team members use daily/weekly huddles and briefs to stay informed, address issues, share unexpected events, and celebrate successes throughout the diagnostic process.	–	3.9 (1.03)	0.82	–
Team members appropriately use all available methods of diagnostic communication (for example, EHR, face-to-face communication).	–	4.0 (0.97)	0.79	–
Communication	*5*	*3.6 (0.78)*	*0.89*	*0.88*
Team members actively exchange information (for example, brief, clear, specific, timely, communication, confirmed by check-backs) that supports effective communication in the diagnostic process.	–	3.7 (0.77)	0.87	–
Team members work collaboratively with other members and access information (for example, EHR) when needed, to inform the diagnostic process.	–	3.9 (0.80)	0.88	–
Team members within our setting hold one another accountable for using structured communication tools (for example, SBAR, call-outs, check-backs, handoff techniques) to facilitate communication.	–	3.4 (1.08)	0.87	–
When communicating with external team specialists, providers and staff consistently use structured referral tools (for example, check-backs, handoff techniques).	–	3.4 (1.02)	0.87	–
Reflective practice (e.g., ask, listen, act) is used consistently in the diagnostic process during interactions (for example, patient–provider, provider–provider, provider–staff).	–	3.4 (0.96)	0.87	–
Leadership	*7*	*3.8 (0.88)*	*0.95*	*0.91*
Leaders ensure that all team members understand the goals and vision for effective communication in the diagnostic process (for example, patient goals, shared model for plan of care) and hold each other accountable (for example, use metrics for tracking improvement, debriefs, huddles).	–	3.7 (0.94)	0.94	–
Leaders provide resources for the diagnostic team to effectively facilitate communication both internally and externally.	–	3.7 (0.96)	0.94	–
Leaders support a balanced workload within the team and delegate tasks consistent with roles and responsibilities of team members.	–	3.6 (1.04)	0.95	–
Leaders act as a liaison for resolving team issues, system issues, and any breakdown in communication.	–	3.8 (1.02)	0.94	–
Leaders set expectations for participation in effective communication practices (for example, briefs, huddles, debriefs) in the diagnostic process.	–	3.8 (1.03)	0.94	–
Leaders reinforce good practices by celebrating diagnostic team successes.	–	3.8 (1.04)	0.94	–
Leaders model teamwork behaviors.	–	3.9 (1.01)	0.94	–
Monitoring	*4*	*3.6 (0.82)*	*0.90*	*0.89*
Team members routinely assess communication practices to identify opportunities for improvement (for example, this survey tool, debriefing events, safety culture surveys).	–	3.5 (0.91)	0.87	–
Team members regularly review systems intended to support the diagnostic process (for example, scheduling, test results, consultations) for gaps and improvement opportunities.	–	3.5 (0.93)	0.87	–
Team members have a systematic process in place to capture and learn from near misses and no-harm adverse events that occur because of communication gaps.	–	3.7 (0.96)	0.89	–
Team members establish goals, share with the diagnostic team, and implement action plans after assessments.	–	3.6 (0.94)	0.86	–

(*continued on next page*)

Table 2. (*continued*)

Dimension	Number of Items*	Mean (SD)	Cronbach's Alpha†	Parallel Forms Reliability
Mutual Support	*4*	*3.5 (0.90)*	*0.89*	*0.89*
Team members are held accountable for proactively assisting each other in the diagnostic process (for example, catching and correcting communication failures, providing task assistance).	–	3.7 (0.95)	0.85	–
Team members freely provide timely and constructive feedback to each other to improve the diagnostic process.	–	3.5 (1.01)	0.84	–
Team members feel safe raising issues, sharing concerns, and advocating for patient needs.	–	3.8 (1.00)	0.89	–
Team members attempt to resolve conflicts using structured communication tools (for example, assertive statements, two-challenge rule, DESC script).	–	3.2 (1.14)	0.88	

* Overall ratings for each subdimension were excluded for this analysis.

† Alpha values for overall scale and dimensions are raw alphas, and values for items are alphas if item deleted from domain subscale.

SD, standard deviation; TAT, Team Assessment Tool; SBAR, Situation, Background, Assessment, Recommendation; EHR, electronic health record; DESC, Describe, Express, Specify, Consequences.

Table 3. Generalizability Study Findings

Source	Variance	% Variance
Respondent	0.54	
Dimension	1.16	59.6
Site	0.02	
Site by Dimension	0.004	0.2
Residual	0.21	

developing action plans. Psychometric design and testing add reliability and validity to the instrument, increase the strength of evaluation findings, and, most importantly, add population-level generalizability with external validity, as results can then be compared within the same setting or with other settings using the same instrument.

The TAT was observed to have excellent internal and parallel forms reliability. These are two fundamental attributes of a high-quality measure. The overall scale and each dimension surpassed heuristic thresholds associated with sufficient reliability for use. A scale's reliability places an upper limit on the relationship that scale can have with other measures (that is, a scale cannot produce a correlation with another measure greater than that scale correlates with itself). The high reliabilities observed provide strong validity evidence for the TAT's use for tracking institutional changes over time, evaluating the impact of an intervention, or comparing TAT scores across different facilities. The G-study findings provide further evidence of the validity of the TAT. The two largest sources of variance in TAT scores were the respondent and the TAT scale dimension. These are both desirable facets of variation, and the large, combined percentage of variance accounted for in these two facets alone (87.6%) provides strong validity evidence for the TAT. Respondent variation is interpreted as systematic differences in TAT scores based on the individual respondent completing the survey. Different respondents view teamwork in the diagnostic process differently. Variation associated with the dimension is interpreted as systematic differences in rating each of the five TAT dimensions. This is desirable, as it indicates that respondents see the TAT dimensions as unique and distinct components of the overall scale. Only 1.2% of variance in TAT scores was associated with the organization, and 0.2% with the interaction between organization and dimension. These are both important dimensions given the TAT's purpose of assessing local teamwork practices and opportunities related to diagnostic safety. Larger percentage of variance associated with the site facet would indicate systematic differences in TAT scores across sites, and the interaction between site and dimension would indicate that people at different sites systematically rated the dimensions differently. As both facets accounted for relatively small amounts of variance, there is limited support for the utility of the TAT in identifying unique local needs at this time. The unexpectedly low percentage of variance accounted for by the site and an interaction between site and dimension facets of variation could be explained by the relatively small number of sites included in these data. Due to attrition at the site level, only seven of the nine recruited sites were included in the final data.

Adhering to the evidence of generating reliable and valid instruments, in combination with the precursor psychometric testing, set the items and domains of the TAT for the best possible outcome, as illustrated by the excellent Cronbach's alpha for the overall instrument (0.97) and high G-study variance on respondents (28.0%) and dimensions (59.6%). Findings demonstrate the value of time spent incorporating methodological rigor into the tool

and iterating the items on the instrument prior to full psychometric evaluation. Reviewers involved in the development of the tool and an external random sample of nonaffiliates of the tool were integral precursors to full psychometric evaluation. The combination of approaches primed the TAT instrument for reliability and validity testing.

Limitations

This study has four limitations. First, rates of careless respondents were high as determined by the longstring index of response invariance, and the Mahalanobis distance index of multivariate outliers (that is, random responses). This is common in organizational surveys with no risk or consequences for respondents and likely amplified by the extended length of the parallel forms survey requiring twice as many responses compared to the stand-alone TAT. By adopting conservative thresholds for eliminating careless respondents, we report a more accurate estimate of the TAT's reliability and generalizability. However, as reported in Appendix 4, results of reliability and generalizability analyses were similar when conducted with the full sample. In practice, administering only the TAT itself (that is, reducing survey items in half) will reduce respondent burden. In addition, this tool is designed to guide local improvement efforts, so there will be more for individual respondents to gain from conscientious responses. However, the lower-than-anticipated response rates introduce uncertainty into this study's findings, which can be resolved best through future validation of the TAT conducted within its intended context of use—implementation of teamwork for diagnostic safety programs. Second, the ability to assess whether the TAT is useful for detecting local (that is, unit- or practice-level) opportunities to improve was limited by the small number of participating organizations as well as variability of units within organizations, particularly for the larger participating organizations. Related measures (such as safety culture) are known to vary more within an organization (by unit) than between.[19] We were not able to capture unit or practice identifiers within the organization, so we were limited to this higher level of analysis. Third, responses within each facility could have come from a wide variety of roles. As the quality of teamwork can be viewed differently by different individuals from different professions,[45] this could increase within facility variation of scores and complicate the detection of between-site differences. Fourth, none of the participating sites had implemented the TeamSTEPPS for Diagnosis Improvement curriculum. As the content of the TAT is rooted in this curriculum, greater differences may be observed across sites as this program is implemented. In addition, the referent shift phenomenon in which respondents change how they respond to items based on their changes in expectations or knowledge after participating in training is common during team training programs[46] and should be explored in use of the tool as a pre- and posttraining evaluation tool. Future research can explore larger samples of organizations and control for role types to better understand the utility of the TAT to assess unique local teamwork strengths and opportunities in the diagnostic process across organizations.

In future work, the TAT has multiple potential uses as a quantifiable measure of diagnostic patient safety teamwork and communication. In practice as a validated instrument, it has the potential to drive local health care setting change, as it measures the maturity phase of key teamwork and communication factors (that is, structure, communication, leadership support, situation monitoring, and mutual support), which are critical domains for effective teamwork, to improve diagnostic clinical patient outcomes. In research, by using this tool, studies can accurately assess and measure the impact of the TeamSTEPPS for Diagnosis Improvement Course on enhancing diagnostic teamwork and communication among health care team members during implementation. In policy, this tool can be part of an assessment bundle that identifies gaps in diagnostic health care delivery, informing policymakers, stakeholders, and health care providers about areas that need development (for example, better strategies for diagnostic patient safety, investing in health care infrastructure, increasing access to team resources) to enhance the quality of care provided to patients. Further, by being valid and reliable, the TAT can be used as part of the TeamSTEPPS for Diagnosis Improvement Course, or separately from the course, as an independent diagnostic teamwork and communication measurement instrument.

CONCLUSION

The results of the psychometric evaluation demonstrate that the TeamSTEPPS for Improving Diagnosis TAT is psychometrically sound. It is a reliable and valid instrument for assessing teamwork and communication among diagnostic teams. The TAT can be used in any health care setting to confidently evaluate teamwork and communication levels among diagnostic teams. Using this tool, health care teams can now benchmark diagnostic teamwork and communication using self-comparators, peer comparators, and trendline comparators to initiate or continue diagnostic improvement activity. The TAT adds an evidence-based, psychometrically sound tool to the repertoire of patient safety assessment instruments and a novel and needed measurement tool to help advance diagnostic teamwork and communication to improve patient care and outcomes.

Funding. This work was funded by the Agency for Healthcare Research and Quality (AHRQ) under contract number HHSP233201500022I/75P00119F37006. This work does not reflect the views of AHRQ, and is the perspective of the authors.
Conflicts of Interest. All authors report not conflicts of interest.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcjq.2023.08.009.

Kisha J. Ali, PhD, MS, is Research Scientist, MedStar Institute for Quality and Safety, MedStar Health Research Institute, Columbia, Maryland. **Christine A. Goeschel, ScD, MPA, RN**, was Vice President, MedStar Institute for Quality and Safety, and Professor, Georgetown University School of Medicine, Washington, DC. **Melissa M. Eckroade, MHA**, is Project Manager, MedStar Institute for Quality and Safety. **Katie N. Carlin, MBA**, is Vice President, Business Development and Growth, American Nurses Association, Silver Spring, Maryland. **Monika Haugstetter, MHA, MSN, RN, CPHQ**, is Program Officer, General Patient Safety Division, Center for Quality Improvement and Patient Safety, Agency for Healthcare Research and Quality, Rockville, Maryland. **Margie Shofer, BSN, MBA**, is Division Director, General Patient Safety Program, Center for Quality Improvement and Patient Safety, Agency for Healthcare Research and Quality. **Michael A. Rosen, PhD, MA**, is Professor, Anesthesiology and Critical Care Medicine, Johns Hopkins Armstrong Institute for Patient Safety and Quality, Johns Hopkins University School of Medicine, Baltimore. Please address correspondence to Kisha J. Ali, kisha.j.ali@medstar.net.

REFERENCES

1. Society to Improve Diagnosis in Medicine. What Is Diagnostic Error? Accessed Sep 7, 2023. https://www.improvediagnosis.org/what-is-diagnostic-error/.
2. National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care, Washington, DC: National Academies Press, 2015. Accessed Sep 8, 2023 https://nap.nationalacademies.org/catalog/21794/improving-diagnosis-in-health-care .
3. Risk Management Foundation of the Harvard Medical Institutions, CRICO Strategies. Malpractice Risks in Communication Failures: 2015 Annual Benchmarking Report. 2015. Accessed Sep 8, 2023. https://www.candello.com/Insights/Candello-Reports/Communications-Report.
4. Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. BMJ Qual Saf. 2014;23:359–372.
5. The Joint Commission. Most Commonly Reviewed Sentinel Event Types. (Updated: Feb 2, 2021.) Accessed Sep 8, 2023. https://www.jointcommission.org/-/media/tjc/documents/resources/patient-safety-topics/sentinel-event/most-frequently-reviewed-event-types-2020.pdf.
6. Rosen MA, et al. An integrative framework for sensor-based measurement of teamwork in healthcare. J Am Med Inform Assoc. 2015;22:11–18.
7. Agency for Healthcare Research and Quality. TeamSTEPPS® 2.0 Team Performance Observation Tool. 2014. Accessed Sep 8, 2023. https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/education/curriculum-tools/teamstepps/instructor/reference/tmpot.pdf.
8. Boateng GO, et al. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018 Jun 11;6:149.
9. Agency for Healthcare Research and Quality. TeamSTEPPS Diagnosis Improvement Course. May 2023. Accessed Sep 8, 2023. https://www.ahrq.gov/teamstepps-program/diagnosis-improvement/index.html.
10. Rashvand F, et al. The assessment of safe nursing care: development and psychometric evaluation. J Nurs Manag. 2017;25:22–36.
11. Brown A, et al. Development and psychometric evaluation of an instrument to measure knowledge, skills, and attitudes towards quality improvement in health professions education: the Beliefs, Attitudes, Skills, and Confidence in Quality Improvement (BASiC-QI) Scale. Perspect Med Educ. 2019;8:167–176.
12. Richard A, Pfeiffer Y, Schwappach DDL. Development and psychometric evaluation of the speaking up about patient safety questionnaire. J Patient Saf. 2021 Oct 1;17:e599–e606.
13. LoBiondo-Wood G, Haber J. Reliability and validity. In: LoBiondo-Wood G, Haber J, editors. Nursing Research. Methods and Critical Appraisal for Evidence-Based Practice. 8th ed. St. Louis: Elsevier Mosby. p. 289–309.
14. Mueller RO, Knapp TR. Reliability and validity. In: Hancock GR, Stapleton LM, Mueller RO, editors. The Reviewer's Guide to Quantitative Methods in the Social Sciences. 2nd ed. New York: Routledge. p. 397–401.
15. Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs. 2015;18:66–67.
16. Knapp TR, Mueller RO. Reliability and validity of instruments. The Reviewer's Guide to Quantitative Methods in the Social Sciences. New York: Routledge. p. 337–341.
17. Agency for Healthcare Research and Quality. TeamSTEPPS® Team Assessment Tool for Improving Diagnosis. AHRQ Publication No. 22-0015. Feb 2022. Accessed Sep 8, 2023. https://www.ahrq.gov/sites/default/files/wysiwyg/teamstepps/diagnosis-improvement/dxsafety-team-assessment-tool.pdf.
18. DiCuccio MH. The relationship between patient safety culture and patient outcomes: a systematic review. J Patient Saf. 2015;11:135–142.
19. Smits M, et al. Measuring patient safety culture: an assessment of the clustering of responses at unit level and hospital level. Qual Saf Health Care. 2009;18:292–296.
20. Jones KJ, et al. The AHRQ hospital survey on patient safety culture: a tool to plan and evaluate patient safety programs Advances in Patient Safety: New Directions and Alternative Approaches, vol 2: Culture and Redesign. Henriksen K, et al., editors, Rockville, MD: Agency for Healthcare Research and Quality, 2008. Accessed Sep 8, 2023 https://www.ncbi.nlm.nih.gov/books/NBK43699/.
21. Rosen MA, et al. Teamwork in healthcare: key discoveries enabling safer, high-quality care. Am Psychol. 2018;73:433–450.
22. Pronovost PJ, et al. Reducing preventable harm: observations on minimizing bloodstream infections. J Health Organ Manag. 2017 Mar 20;31:2–9.
23. Pham JC, et al. CLABSI conversations: lessons from peer–to-peer assessments to reduce central line–associated bloodstream infections. Qual Manag Health Care. 2016;25:67–78.
24. Stone AB, et al. Barriers to and facilitators of implementing enhanced recovery pathways using an implementation framework: a systematic review. JAMA Surg. 2018 Mar 1;153:270–279.
25. Dodge LE, et al. Long-term effects of teamwork training on communication and teamwork climate in ambulatory reproductive health care. J Healthc Risk Manag. 2021;40(4):8–15.
26. Newman-Toker DE. A unified conceptual model for diagnostic errors: underdiagnosis, overdiagnosis, and misdiagnosis. Diagnosis (Berl). 2014;1:43–48.
27. Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. Psychol Bull. 1988;103:265–275.

28. Jeon Y, et al. Developing and psychometric testing of the anaesthesia nursing competence scale. J Eval Clin Pract. 2020;26:866–878.
29. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill, 1994.
30. Desmedt M, et al. Systematic psychometric review of self--reported instruments to assess patient safety culture in primary care. J Adv Nurs. 2018;74:539–549.
31. Wu Q, et al. Development and psychometric evaluation of the Patient Engagement in Health Care Questionnaire. J Nurs Care Qual. 2020;35:E35–E40.
32. Wilson M, et al. Psychometric evaluation of the Creighton Competency Evaluation Instrument in a population of working nurses. J Nurs Meas. 2022 Mar 1;30:148–167.
33. Hibbard JH, et al. Development and testing of a short form of the Patient Activation Measure. Health Serv Res. 2005;40:1918–1930.
34. Ward MK, Meade AW. Dealing with careless responding in survey data: prevention, identification, and recommended best practices. Annu Rev Psychol. 2023 Jan 18;74: 577–596.
35. Arthur W Jr, Hagen E, George F Jr. The lazy or dishonest respondent: detection and prevention. Annual Review of Organizational Psychology and Organizational Behavior. 2021;8:105–137.
36. Yentes RD. In Search of Best Practices for the Identification and Removal of Careless Responders. North Carolina State University, 2020.
37. Yentes RD, Wilhelm F. Careless: Procedure for Computing Indices of Careless Responding, version 1.2.1 (computer software), 2021.
38. de Vet HCW, et al. Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. J Clin Epidemiol. 2017;85: 45–49.
39. McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires. 3rd ed. New York: Oxford University Press, 2006.
40. Kraiger K, Teachout MS. Generalizability theory as construct-related evidence of the validity of job performance ratings. Hum Perform. 1990;3:19–35.
41. Crossley J, et al. Generalisability: a key to unlock professional assessment. Med Educ. 2002;36:972–978.
42. Crossley J, et al. 'I'm pickin' up good regressions': the governance of generalisability analyses. Med Educ. 2007;41:926–934.
43. R Foundation. The R Project for Statistical Computing. Accessed Sep 8, 2023. http://www.R-project.org/.
44. Moore CT. G Theory: Apply Generalizability Theory with R. R package version 0.1, Oct 30, 2016. Accessed Sep 8, 2023 https://CRAN.R-project.org/package=gtheory.
45. House S, Havens D. Nurses' and physicians' perceptions of nurse-physician collaboration: a systematic review. J Nurs Adm. 2017;47:165–171.
46. Rosen MA, et al. How can team performance be measured, assessed, and diagnosed?. In: Salas E, Frush K, editors. Improving Patient Safety Through Teamwork and Team Training. New York: Oxford University Press. p. 59–79.