Assessing Health Care Team Performance: A Review of Tools and the Evidence Supporting Their Use

Shannon L. Marlow, M.S., Christina Lacerenza, M.S., Chelsea Iwig, M.S.,

& Eduardo Salas, Ph.D.

Rice University

Eduardo Salas, Ph.D.
Rice University
Department of Psychology-MS 25
Sewall Hall 429C
6100 Main St.
Houston, TX 77005-1827
(713) 348-3917
Eduardo.salas@rice.edu

Assessing Health Care Team Performance

Overview

Health care practitioners and researchers alike are increasingly recognizing the role of teamwork in ensuring effective patient care and safety, as reflected by the increased implementation of health care team training in health care organizations (Beach, 2013; Hughes et al., in press; Weaver et al., 2010). Supporting the utility of this approach, a recent meta-analysis found that health care team training was linked to a host of positive outcomes within the health care context, including reduced patient mortality, reduced medical error, and improved teamwork on-the-job (Hughes et al., in press). Given these findings, there is a clear case for health care organizations to emphasize teamwork and health care team training as viable approaches to enhancing patient care. One critical component of ensuring the success of both of these initiatives is accurately measuring health care team performance. Confirming this need, countless health care team performance measures have been developed to evaluate teamwork in the medical context (Jeffcott & Mackenzie, 2008).

However, to ensure accurate measurement and mitigate inaccurate conclusions, certain steps must be completed and data collected before a scale can be implemented with the sample of choice (Cronbach & Meehl, 1955; DeVellis, 2012). Specifically, researchers caution that robust validation and reliability data should be gathered before promoting the use of a measure (DeVellis, 2012; Guion, 1980). Without collecting this data, the accuracy of a measure cannot be assumed. Despite existing evidence for the utility of validating measures before use (e.g., DeVellis, 2012), not all measures have been developed in accordance with recommended guidelines. This concern is especially applicable to industry, given the often urgent and constrictive restraints associated with collecting any given set of data. In regards to the health

care industry, where interest in teamwork and, consequently, measures of team performance is growing at a frequent rate, it is especially important to ensure that measures are first validated before being implemented. Thus, the aim of the current effort is to conduct a systematic review on health care team performance measures and organize them by this data.

Objectives

The current effort seeks to systematically organize current health care team performance measures by:

- Identifying existing health care team performance measures
- Synthesizing existing evidence supporting the validity and reliability for each measure
- Presenting additional information pertaining to the use and implementation of the measures (e.g., ease of administration)

By completing this review, we ultimately seek to provide health care practitioners comprehensive information about current measures available to evaluate health care team performance as well as evidence supporting their use. We ultimately hope this information can serve as a guide for a choice of a health care team performance measure given current needs and goals.

Method

Literature Search

We conducted a literature search of the following databases: Academic Search Premiere, CINAHIL, Google Scholar, MEDLINE, Ovid, PubMed, PsychInfo, and Science Direct. We leveraged various combinations of the following search terms: *team, performance, health care,*

and *medical.* We included both published and unpublished articles that detailed information about performance measures intended for use in the health care context. Measures were excluded if they did not specifically pertain to *team* performance or if they were not intended for use with a health care team sample. We only included measures if they were not originally intended for use with a health care team sample but were implemented with such a sample at some point. 53 measures were identified with this strategy. We also attempted to identify any additional articles incorporating or testing the measures in order to collect any evidence for validity and/or reliability that is currently available.

Coding

Coders independently extracted information from the measures. The coders were three doctoral students with expertise in the domains of teams and performance. Each article was coded by at least two of these three individuals. Agreement between the coders was 86% and any discrepancies were resolved through discussion. We extracted information relevant to the following broad categories: (1) general characteristics of measures (i.e., accessibility; clarity of language; instrument type; applicability; objectivity vs. bias), (2) validity (i.e., criterion validity; construct validity; content validity), and (3) reliability (i.e., inter-rater/inter-observer reliability; internal consistency; test-retest reliability). Appendix A presents information pertaining to the general characteristics of each measures and Appendix B presents information for the reliability and validity of each measure.

Accessibility. Accessibility reflects how readily accessible the instrument is to a layperson. This was coded using the following categories: (1) open access (i.e., the measure is freely available and accessible online), (2) subscription required (i.e., the measure is published in a journal article that is available via a journal subscription), (3) copyrighted (i.e., the measure is

copyrighted and cannot be utilized unless the appropriate permission is attained), and (4) unpublished (i.e., the measure is unavailable online or elsewhere but the use of the measure was documented elsewhere, although the measure itself is not published).

Clarity of Language. Clarity of language refers to how easily understood a measure is to someone with no background experience. The following categories were developed for this category: (1) high (i.e., the measure uses no jargon and can be readily understood by someone with no relevant background experience), (2) moderate (i.e., the measure uses some jargon but can still be largely understood by someone with no relevant background experience), and (3) low (i.e., the measure uses a high amount of jargon and can only be understood by someone with relevant background experience).

Instrument Type. Although some instruments can be implemented by incorporating observers *and* via self-report, or by using multiple methods, we created this category to reflect the method by which the instrument was originally intended to be used. In other words, this coding category reflects the manner in which the instrument was originally developed and validated with. The following categories were developed: (1) self-report (i.e., the measure is intended to be completed by someone rating their team members and themselves) and (2) observer (i.e., the measure is intended to be completed by someone apart from the team).

Applicability. Applicability is defined as the degree to which the measure can be readily implemented with a generic team. We developed the following categories to code for this element: (1) generic (i.e., the measure was developed for team assessment in general and does not include any specific references to health care teams), (2) adaptable (i.e., the measure contains terms specific to a profession but could be easily adapted to apply to teams in additional health

care professions), and (3) focused (i.e., the measure was created for use in a specific context and cannot be adapted to apply to additional teams without significant revision).

Objectivity vs. Bias. Researchers argue that measures should be grounded in observable behaviors to reduce the amount of subjectivity required, and thus bias, in rating performance (e.g., Rosen et al., 2008). To address this measurement concern, we created this category to reflect the degree to which a measure appears to be tied to observable behaviors. The following categories were developed: (1) very objective (i.e., the majority of items in a measure are quantifiable, measureable markers of behaviors such as 'the team member verified that information was understood during communication'; the measure is intended for use with an observer), (2) fairly objective (i.e., the majority of items in a measure are less quantifiable, measureable markers of behaviors and may incorporate vaguer items such as 'the team worked together effectively'; the measure is intended for use with an observer), and (3) less objective (i.e., the measures are self-report). We note that self-report measures often produce inflated scores as individuals generally rate themselves more highly than an observer would (e.g., Blume, Ford, Baldwin, & Huang, 2010) and consequently rated all self-report measures as less objective.

Criterion Validity. Criterion validity reflects the extent to which a measure is related to an external outcome and was founded with the underlying idea that *concurrent* and *predictive* validity are two facets of this construct (Guion, 2011). Concurrent validity refers to a relationship between the measure and an outcome when the data for both sources is collected at the same time (APA, 1974). Predictive validity also encompasses a relationship between the measure and an outcome but refers to a context where data for the outcome is collected at a later time (APA, 1974). In general, concurrent validity is considered some evidence for criterion validity but not considered a substitute for predictive validity (APA, 1974). In line with this reasoning, we

created the following categories to assess the evidence available for criterion validity of each measure: (1) strong (i.e., there is evidence of predictive validity or there is evidence for both predictive and concurrent validity; for example, data for a theoretically related outcome is collected several months after the health care team performance measure and these two measures exhibit a moderate correlation), (2) some (i.e., there is evidence of concurrent validity; for example, data for the health care team performance measure and a theoretically related outcome measure are collected at the same time and exhibit a moderate correlation), and (3) unable to find supporting evidence (i.e., we were unable to identify or find any evidence for either concurrent or predictive validity).

Construct Validity. Construct validity refers to "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231). There is currently no universally agreed upon method for evaluating construct validity, however, most researchers suggest that an empirical method should be used to evaluate this element and the more methods used, the more evidence for construct validity (Brown, 2000). For example, Brown (2000) notes that any of the following methods could be used to establish construct validity: "content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pre-test-posttest intervention studies, multi-trait/multi-method studies, etc" (p. 10). Thus, we developed our coding scheme to assess whether any of the above, or related, methods were employed to determine whether the measure was exhibiting evidence in line with what theory surrounding the construct would suggest.

Moreover, we also considered any available evidence for *convergent* and *discriminant* validity. These two components are considered two facets of construct validity (Campbell & Fiske, 1959). Convergent validity refers to the degree to which two measures that should

theoretically be related are related whereas discriminant validity is conceptualized as the extent to which a measure is *unrelated* to a measure that it should, theoretically, not be related to (Campbell & Fisk, 1959). We note that researchers caution that both convergent validity and discriminate validity are necessary to support construct validity (Trochim & Donnelly, 2006) and took this notion into account when developing our coding scheme.

Although multiple methods of testing are preferable, we realize the practical constraints associated with gathering enough data to implement multiple testing techniques. Consequently, we considered any empirical testing that produced results in accordance with theory surrounding team performance evidence of strong construct validity. Specifically, we used the following coding scheme: (1) strong evidence (i.e., at least two empirical techniques are used to assess the measure and produce evidence in line with theorized results *or* there is evidence for both convergent and discriminant validity), (2) some evidence (i.e., one empirical technique is used to assess the measure and produces limited or strong evidence in line with theorized results *or* there is evidence for discriminant validity *or* there is evidence for convergent validity), and (3) unable to find supporting evidence (i.e., we were unable to identify or find any evidence for construct validity).

Content Validity. Content validity is defined as "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose" (Haynes, Richard, & Kubany, 1995, p. 238). Researchers suggest that this can be addressed by ensuring that steps are taken to include every domain of the construct being assessed. Some possible methods of evaluating the content validity of a measure include piloting the items with a relevant sample, gaining consensus from experts, conducting an extensive literature review, synthesizing existing measures that have been previously assessed for content

validity, and basing items on observed behavior. Although, ideally, all measures will have undergone a revision process that incorporates expert review (Smith & McCarthy, 1995) we note that practical constraints may limit the methods researchers are able to utilize. A review of the available measures suggests that this approach is rarely used. Consequently, we developed our coding scheme to reflect this, using the following codes: (1) detailed information available (i.e., at least more than one of the methods described above is used to ensure content validity), (2) some information available (i.e., only one of the methods described above is used to ensure content validity), and (3) unable to find supporting evidence (i.e., we were unable to identify or find any evidence for content validity).

Inter-Rater Reliability. Inter-rater reliability, also known as inter-observer reliability, only applies to measures that are intended to be utilized with an observer. This construct refers to the extent to which different observers consistently rate the same behaviors (Guion, 2011; James, Demaree, & Wolf, 1984). In other words, inter-rater reliability reflects whether raters are scoring behaviors in roughly the same manner. Inter-rater reliability can be assessed using the $r_{wg(j)}$ statistic (James et al., 1984) and the intraclass correlation coefficient (ICC) but there are a host of additional methods by which this can be evaluated. Ultimately, however, the metric utilized must assess the degree of consistency between the scores of the different observers. Thus, to rate this category we used the following scheme, interpreting the statistic used with guidelines available specific to that particular metric: (1) high (i.e., the statistic utilized to assess inter-rater reliability reflected a high degree of consistency among raters), (2) moderate (i.e., the statistic utilized to assess inter-rater reliability reflected a moderate degree of consistency among raters), (3) low (i.e., the statistic utilized to assess inter-rater reliability reflected a low degree of consistency among raters), and (4) N/A (i.e., the measure was intended to be administered via self-report).

Internal Consistency. The internal consistency of a measure refers to the extent to which the test items of a measure consistently reflect the intended characteristic (Guion, 2011). The most common method of evaluating internal consistency is via assessing Cronbach's coefficient alpha (1951). Therefore, we utilized Cronbach's alpha data, if available, to assess the internal consistency of the measures. We interpreted Cronbach's alpha in accordance with preexisting guidelines (e.g., Nunnally, 1978) by applying the following coding scheme: (1) high (i.e., the Cronbach's alpha of the measure ranged from .7 to higher), (2) moderate (i.e., the Cronbach's alpha of the measure ranged from .6 to higher), (3) low (i.e., the range of the Cronbach's alpha of the measure included a score lower than .6), and (4) unable to find supporting evidence (i.e., we were unable to identify or find any evidence for internal consistency).

Test-Retest Reliability. Test-retest reliability, or repeatability, is assessed by collecting data with one measure at two separate time points. Specifically, the measure must be administered under the same conditions to the same sample (Portney & Watkins, 2000). The scores from the two different testing periods can subsequently be related in some manner, typically via correlation, to determine the extent to which the scores produced from the measure are the same over time. We used the following categories to assess test-retest reliability: (1) high (i.e., the statistic utilized to assess test-retest reliability reflected a high degree of reliability), (2) moderate (i.e., the statistic utilized to assess test-retest reliability reflected a moderate degree of reliability), (3) low (i.e., the statistic utilized to assess test-retest reliability reflected a low degree of reliability), and (4) unable to find supporting evidence (i.e., we were unable to identify or find any evidence for test-retest reliability). Note that, as with inter-rater reliability, we interpreted the strength of the statistic in accordance with available guidelines.

Results

We organize results in the following categories: overall (i.e., information pertaining to general characteristics about the measures), validity (i.e., information regarding criterion, construct, and content validity of the measures), and reliability (i.e., information related to inter-rater reliability, internal reliability, and test-retest reliability of the measures). All percentages are calculated from the total number of measures ($k = 53$) unless otherwise stated.

General Characteristics

Information pertaining to the general characteristics of measures is summarized in Table 1.

Table 1
General Characteristics of Measures

Characteristic	Number of Articles (Percentage)
Availability	
Open Access	11 (20.8%)
Subscription Required	24 (45.3%)
Copyrighted	15 (28.3%)
Unpublished	3 (5.7%)
Clarity of Language	
High	48 (90.6%)
Moderate	4 (7.5%)
Low	1 (1.9%)
Type of Instrument	
Self-report	34 (64.2%)
Observer	19 (35.8%)
Applicability	
Generic	22 (41.5%)
Adaptable	22 (41.5%)
Focused	9 (17%)
Objectivity vs. Bias	
Very Objective	35 (66%)
Fairly Objective	12 (22.6%)
Less Objective	6 (11.3%)

Accessibility. A total of 11 (20.8%) measures were open access; 24 (45.3%) measures were available through journal subscription; 15 (28.3%) measures were copyrighted; and 3 (5.7%) were unavailable because they were unpublished.

Clarity of Language. The majority of measures used language that was high in clarity (*k* = 48, 90.6%), however, 4 (7.5%) measures used language that was moderate in clarity. Only 1 (1.9%) measure used language that was low in clarity.

Instrument Type. Measures were created either with the intention of being implemented with self-report (*k* = 34, 64.2%) or with an observer (*k* = 19, 35.8%).

Applicability. The majority of measures were generic (*k* = 22, 41.5%) or adaptable (*k* = 22, 41.5%). An additional 9 (17%) measures were more focused in nature.

Objectivity vs. Bias. As a high number of measures were intended to be administered via self-report, there was a corresponding high amount of measures that were less objective (*k* = 35, 66%). However, we found 12 (22.6%) measures that were fairly objective and 6 (11.3%) that were very objective.

Validity

Validity information pertaining to the measures is summarized in Table 2.

Table 2
Summary of Available Validity Information related to Measures

Characteristic	Number of Articles (Percentage)
Criterion Validity	
High evidence	1 (1.9%)
Some evidence	16 (30.2%)
No evidence identified	36 (67.9%)
Construct Validity	
High evidence	20 (37.7%)
Some evidence	14 (26.4%)
No evidence identified	19 (35.8)
Content Validity	
Detailed information available	39 (73.6%)
Some information available	10 (18.9%)

No information available	4 (7.5%)

Criterion Validity. Only 1 (1.9%) measure had strong criterion validity evidence associated with it. An additional 16 (30.2%) measures had some evidence for criterion validity but the majority of measures (k = 36, 67.9%) had no evidence supporting criterion validity whatsoever, requiring additional testing.

Construct Validity. Approximately half of the identified measures had strong evidence supporting their construct validity (20, 37.7%). An additional 14 (26.4%) measures had some evidence supporting construct validity, however, many measures (k = 19, 35.8%) had no evidence associated with them.

Content Validity. Many measures had detailed information available regarding how content validity was established or considered (k = 39, 73.6%). An additional 10 (18.9%) measures had some information related to content validity available and only 4 (7.5%) measures had no information available whatsoever.

Reliability

Reliability information about the measures is presented in Table 3.

Table 3
Summary of Available Reliability Information related to Measures

Characteristic	Number of Articles (Percentage)
Inter-Rater Reliability	
High	4 (21.1%)
Moderate	4 (21.1%)
Low	6 (31.6%)
No information identified	5 (26.3%)
N/A	34 (64.2%)
Internal Consistency	
High	27 (50.9%)
Moderate	5 (9.4%)
Low	4 (7.5%)
No information identified	17 (32.1%)
Test-retest Reliability	

High	4 (7.5%)
Moderate	4 (7.5%)
Low	3 (5.7%)
No information identified	42 (79.2%)

Inter-Rater Reliability. It is important to note that most measures were intended to be completed via self-report (k = 34, 64.2%) and that inter-rater reliability is inappropriate to consider in these cases. Thus, this category reflects only the 19 measures intended to be completed with observers. Of these 19 measures, 4 (21.1%) exhibited a high degree of inter-rater reliability, 4 (21.1%) a moderate degree, and 6 (31.6%) a low degree. Finally, 5 (26.3%) measures had no information related to inter-rater reliability associated with them.

Internal Consistency. The majority of measures (k = 27, 50.9%) demonstrated high internal consistency. Only 5 (9.4%) measures exhibited moderate internal consistency and only 4 (7.5%) demonstrated low internal consistency. However, there were 17 (32.1%) measures for which there was no information related to internal consistency available.

Test-Retest Reliability. It was uncommon for measures to provide information related to test-retest reliability (k = 42, 79.2%). There were only 4 (7.5%) measures that demonstrated a high degree of test-retest reliability. An additional 4 (7.5%) measures had moderate test-retest reliability. Finally, 3 (5.7%) measures exhibited a low degree of test-retest reliability.

Summary

A total of 53 medical team performance measures were identified through our systematic literature search. Broadly, we categorized them based on general characteristics, reliability information, and validity information. Below, we elaborate on the trends evident in each overall category.

General Characteristics.

- The majority of measures were available through journal subscription
- A small subset of articles were freely available
- Most measures use a high clarity of language
- Very few measures use jargon
    - Clarity of language is not generally a concern
- The majority of measures are intended for self-report use
    - As these measures are easily administered, most identified measures are fairly easy to implement and require little, if no, training
- Observer measures were generally fairly objective
    - Very objective observer measures were less common
- Most measures were generic or adaptable and applicable to most health care teams
    - Very few measures were created for use with specific teams

Validity.

- Only 1 article had strong evidence for criterion validity
- The majority of measures had no evidence for criterion validity
    - Whether measures are actually related to constructs they should theoretically predict may be a general concern/limitation
- Most measures had some or strong construct validity evidence
- However, some measures had no construct validity evidence
    - These measures should be used with caution
    - They may not be measuring what they claim to measure
    - Results may not be accurate
    - Validation studies are required
- In general, there was detailed information available supporting content validity
- Only a small subset of articles were lacking information about content validity

Reliability.

- Low inter-rater reliability was common
    - When implementing an observer measure, this may be a concern
    - Steps should be taken to ensure inter-rater reliability is consistent
    - Rater training may be one method of addressing this concern
- Internal consistency was generally high
    - Some measures had no information available related to internal consistency
    - Measures without internal consistency data should be tested for consistency before being implemented

        - Without ensuring reliability, measures may produce distorted, ineffective results
- Few measures provided information about test-retest reliability
    - It may be helpful, when implementing a measure, to collect this data if it has not already been assessed

Conclusion

We identified 53 measures intended for use with health care teams to measure team performance and categorized detailed information about each measure. In general, they were easily implemented with a new sample, as clarity of language was generally high and the measures were mostly intended for administration via self-report. There were also a large number of measures validated for use with observers, which may be preferable given the goals of measurement and to avoid self-report biases. However, the objectivity of these measures may be a concern, as it is preferable measures are tied to highly observable behaviors, especially if a high degree of inter-rater reliability has not been established; otherwise, rating may be difficult for observers and lead to inaccurate results. The majority of measures had been assessed for reliability and validity in some manner. However, there was a large amount of measures that had not undergone any validation or reliability testing. When implementing such measures, steps should be taken to ensure that validity and reliability are supported, otherwise results may be inaccurate. Ultimately, the goals of measurement should guide the choice of a measure and the information presented in this report may provide guidance in this respect by presenting detailed information pertaining to characteristics of each health care team performance measure.

References

**denote articles that include measures*

*Agency for Healthcare Research and Quality (AHRQ). (2012, December). Team Assessment Questionnaire (TAQ). Rockville, MD: AHRQ. Retrieved from http://www.ahrq.gov/professionals/education/curriculum-tools/teamstepps/instructor/reference/tmassess.html

*AHRQ. (2014, March). Teamwork Perceptions Questionnaire (T-TPQ). Rockville, MD: AHRQ. Retrieved from http://www.ahrq.gov/professionals/education/curriculum-tools/teamstepps/longtermcare/sitetools/tmpot.html

*AHRQ. (2014, October). Team Performance Observation Tool (TPOT). Rockville, MD: AHRQ. Retrieved from http://www.ahrq.gov/professionals/education/curriculum-tools/teamstepps/longtermcare/sitetools/tmpot.html

American Psychological Association (APA). (1974). *Standards for educational and psychological tests.* Washington, DC: APA.

*Anderson, N. R., & West, M. A. (1998). Measuring climate for work group innovation: Development and validation of the team climate inventory. *Journal of Organizational Behavior, 19*(3), 235-258.

*Archibald, D., Trumpower, D, & MacDonald, C.J. (2014) Validation of the interprofessional collaborative competency attainment survey (ICCAS). *Journal of Interprofessional Care, 28*(6) 553-558.

*Baggs, J. G. (1994). Development of an instrument to measure collaboration and satisfaction about care decisions. *Journal of Advanced Nursing, 20,* 176-182.

*Bailey, D. B., Helsel-DeWert, M., Thiele, J. E., & Ware, W. B. (1983). Measuring individual participation on the interdisciplinary team. *American Journal of Mental Deficiencies, 88*, 247-254.

*Batorowicz, B., & Shepherd, T. A. (2008). Measuring the quality of transdisciplinary teams. *Journal of Interprofessional Care, 22*(6), 612-620.

Beach, S. (2013). Annual medical school graduation survey shows gains in teams. Retrieved from https://www.highbeam.com/doc/1G1-338544896.html

Beebe, P., Bawel-Brinkley, K., & O'Leary-Kelley, C. (2012). Observed and self-perceived teamwork in a rapid response team. *Journal for Nurses in Professional Development*, *28*(4), 191-197.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065-1105.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (2000). What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(2), 8-12.

Campell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Capella, J., Smith, S., Philp, A., Putnam, T., Gilbert, C., Fry, W., ... & Ranson, S. (2010). Teamwork training improves the clinical care of trauma patients. *Journal of Surgical Education*, *67*(6), 439-443.

Cooper, S. J., & Cant, R. P. (2014). Measuring non-technical skills of medical emergency teams: An update on the validity and reliability of the Team Emergency Assessment Measure. *Resuscitation*, *85*(1), 31-33.

*Cooper, S., Cant, R., Porter, J., Sellick, K., Somers, G., Kinsman, L., & Nestel, D. (2010). Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM).*Resuscitation*, *81*(4), 446-452.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

DeVellis, R. F. (2012). *Scale development: Theory and applications.* Thousand Oaks, CA: Sage Publications.

*Dyer, W.G. (1987). *Team building*. In S. L. Phillips & R. L. Elledge (Eds.), The team-building source book. San Diego, CA.: University Associates.

Farrell, B., Pottie, K., Woodend, K., Yao, V., Dolovich, L., Kennie, N., & Sellors, C. (2010). Shifts in expectations: evaluating physicians' perceptions as pharmacists become integrated into family practice. *Journal of Interprofessional Care*, *24*(1), 80-89.

*Farrell, B., Pottie, K., Woodend, K., Yao, V. H., Kennie, N., Sellors, C., ... & Dolovich, L. (2008). Developing a tool to measure contributions to medication-related processes in family practice. *Journal of Interprofessional Care*, *22*(1), 17-29.

*Farrell, M. P., Schmitt, M. H., Heinemann, G. D., & Roghmann, K. J. (2001). The Team Anomie Scale: An indicator of poor functioning in health care teams. Unpublished document.

*Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' Non-Technical Skills (ANTS): Evaluation of a behavioural marker system. *British Journal of Anaesthesia, 90*(5), 580-588.

Guion, R. M. (1980). *On trinitarian doctrines of validity. Professional Psychology, 11, 385–398.*

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions.* New York, NY: Routledge.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247.

*Healey, A. N., Undre, S., & Vincent, C. A. (2004). Developing observational measures of performance in surgical teams. *Quality and Safety in Health Care, 13*(suppl 1), i33-i40.

*Heinemann, G. D., Schmitt, M. H., Farrell, M. P., & Brallier, S. A. (1999). Development of an attitudes toward health care teams scale. *Evaluation & the Health Professions, 22*(1), 123-142.

Heinemann, G. D., & Zeiss, A. M. (2002). *Team performance in health care: Assessment and development.* New York, NY: Kluwer Academic/Plenum Publishers.

*Hepburn, K., Tsukuda, R. A., & Fasser, C. (1998). Team skills scale, 1996. In E. L. Siegler, K. Hyer, T. Fulmer, & M. Mezey (Eds.), *Geriatric interdisciplinary team training* (pp. 264-265). New York, NY: Springer Publishing Company.

*Hojat, M., Fields, S. K., Veloski, J. J., Griffiths, M., Cohen, M. J., & Plumb, J. D. (1999). Psychometric properties of an attitude scale measuring physician-nurse collaboration. *Evaluation & the Health Professions, 22*(2), 208-220.

Hojat, M., Gonnella, J. S., Nasca, T. J., Fields, S. K., Cicchetti, A., Scalzo, A. L., ... & Liva, C. (2003). Comparisons of American, Israeli, Italian and Mexican physicians and nurses on the total and factor scores of the Jefferson scale of attitudes toward physician–nurse collaborative relationships. *International journal of nursing studies*, *40*(4), 427-435.

Hojat, M., Nasca, T. J., Cohen, M. J., Fields, S. K., Rattner, S. L., Griffiths, M., ... & Garcia, A. (2001). Attitudes toward physician-nurse collaboration: A cross-cultural study of male and female physicians and nurses in the United States and Mexico. *Nursing Research*, *50*(2), 123-128.

Hughes AM, Gregory ME, Joseph DL, Sonesh SC, Marlow SL, Lacerenza CN, et al. (In press). Saving lives: A meta-analysis on team training in healthcare. *Journal of Applied Psychology.*

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85-98.

Jeffcott, S. A., & Mackenzie, C. F. (2008). Measuring team performance in healthcare: review of research and implications for patient safety. *Journal of Critical Care*, *23*(2), 188-196.

Keebler, J. R., Dietz, A. S., Lazzara, E. H., Benishek, L. E., Almeida, S. A., Toor, P. A., ... & Salas, E. (2014). Validation of a teamwork perceptions measure to increase patient safety. *BMJ Quality & Safety, 23*, 718-726.

*Kenaszchuk, C., Reeves, S., Nicholas, D., & Zwarenstein, M. (2010). Validity and reliability of a multiple-group measurement scale for interprofessional collaboration. *BMC Health Services Research, 10*(1), 1-15.

*Kiesewetter, J., & Fischer, M. R. (2015). The teamwork assessment scale: A novel instrument to assess quality of undergraduate medical students' teamwork using the example of simulation-based ward-rounds. *GMS Zeitschrift für Medizinische Ausbildung, 32*(2), 1-18.

*King, G., Shaw, L., Orchard, C. A., & Miller, S. (2010). The interprofessional socialization and valuing scale: A tool for evaluating the shift toward collaborative care approaches in health care settings. *Work*, *35*(1), 77-85.

*Lamb, B. W., Wong, H. W., Vincent, C., Green, J. S., & Sevdalis, N. (2011). Teamwork and team performance in multidisciplinary cancer teams: Development and evaluation of an observational assessment tool. *BMJ Quality & Safety*, *20,* 849-856.

*Lazar, R. G. (1971). *Team Excellence Questionnaire*. Lawrenceville, GA: Management Skills International.

*Lazar, R. G. (1985). *Personal Productivity and Excellence Potential.* Lawrenceville, GA: Management Skills International.

*Lichtenstein, R., Alexander, J. A., Jinnett, K., & Ullman, E. (1997). Embedded intergroup relations in interdisciplinary team's effects on perceptions of level of team integration. *The Journal of Applied Behavioral Science, 33*(4), 413-434.

Lie, D., May, W., Richter-Lagha, R., Forest, C., Banzali, Y., & Lohenry, K. (2015). Adapting the McMaster-Ottawa scale and developing behavioral anchors for assessing performance in an interprofessional Team Observed Structured Clinical Encounter. *Medical Education Online, 20*, 1-10.

Littlepage, G. E., Cowart, L., & Kerr, B. (1989). Relationships between group environment scales and group performance and cohesion. *Small Group Research*, *20*(1), 50-61.

*Lyk-Jensen, H. T., Jepsen, R. M. H. G., Spanager, L., Dieckmann, P., & Østergaard, D. (2014). Assessing nurse anaesthetists' non-technical skills in the operating room. *Acta Anaesthesiologica Scandinavica, 58*(7), 794-801.

*Malec, J. F., Torsher, L. C., Dunn, W. F., Wiegmann, D. A., Arnold, J. J., Brown, D. A., & Phatak, V. (2007). The mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simulation in Healthcare, 2*(1), 4-10.

*McClane, W.E. (1992). Evaluation and accountability. In American Congress of Rehabilitation Medicine, *Guide to interdisciplinary practice in rehabilitation settings* (pp. 158-172). Skokie, IL: Author.

McCulloch, P., Mishra, A., Handa, A., Dale, T., Hirst, G., & Catchpole, K. (2009). The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. *Quality and Safety in Health Care*, *18*(2), 109-115.

McKay, A., Walker, S. T., Brett, S. J., Vincent, C., & Sevdalis, N. (2012). Team performance in resuscitation teams: comparison and critique of two recently developed scoring tools. *Resuscitation*, *83*(12), 1478-1483.

*Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality and Safety in Health Care, 18*(2), 104-108.

*Moos, R. H. (1986). *Group Environment Scale: Development, Applications, Research* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press, Inc.

Moos, R. H. (1994). *Work environment scale manual*. Consulting Psychologists Press.

Morgan, L., New, S., Robertson, E., Collins, G., Rivero-Arias, O., Catchpole, K., ... & McCulloch, P. (2015). Effectiveness of facilitated introduction of a standard operating procedure into routine processes in the operating theatre: a controlled interrupted time series. *BMJ Quality & Safety*, *24*(2), 120-127.

*National Patient Safety Agency. (2006). The Team Climate Assessment Measurement (TCAM) questionnaire. Aston Organization Development LTD.

*Norris, J., Carpenter, J. G., Eaton, J., Guo, J. W., Lassche, M., Pett, M. A., & Blumenthal, D. K. (2015). The development and validation of the interprofessional attitudes scale:

Assessing the interprofessional attitudes of students in the health professions. *Academic Medicine, 90*(10), 1394-1400.

Nunnally, J. (1978). *Psychometric methods.* New York, NY: McGraw-Hill.

*Ødegård, A. (2006). Exploring perceptions of interprofessional collaboration in child mental health care. *International Journal of Integrated Care, 6,* 1-13.

Ødegård, A., Hagtvet, K. A., & Bjørkly, S. (2008). Applying aspects of generalizability theory in preliminary validation of the Multifacet Interprofessional Collaboration Model (PINCOM). *International Journal of Integrated Care, 8,* 1-11.

*Orchard, C. A., King, G. A., Khalili, H., & Bezzina, M. B. (2012). Assessment of interprofessional team collaboration scale (AITCS): development and testing of the instrument. *Journal of Continuing Education in the Health Professions, 32*(1), 58-67.

*Ottestad, E., Boulet, J. R., & Lighthall, G. K. (2007). Evaluating the management of septic shock using patient simulation. *Critical Care Medicine, 35*(3), 769-775.

*Parsell, G., & Bligh, J. (1999). The development of a questionnaire to assess the readiness of health care students for interprofessional learning (RIPLS).*Medical education*, *33*(2), 95-100.

*Pollard, K. C., Miers, M. E., & Gilchrist, M. (2004). Collaborative learning for collaborative working? Initial findings from a longitudinal study of health and social care students. *Health & Social Care in the Community*, *12*(4), 346-358.

Pollard, K., Miers, M. E., & Gilchrist, M. (2005). Second year scepticism: Pre-qualifying health and social care students' midpoint self-assessment, attitudes and perceptions concerning

interprofessional learning and working. *Journal of Interprofessional Care*, *19*(3), 251-268.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Prentice Hall.

Reid, R., Bruce, D., Allstaff, K., & McLernon, D. (2006). Validating the Readiness for Interprofessional Learning Scale (RIPLS) in the postgraduate context: are health care professionals ready for IPL?. *Medical education*,*40*(5), 415-422.

*Robertson, E. R., Hadi, M., Morgan, L. J., Pickering, S. P., Collins, G., New, S., ... & Catchpole, K. C. (2014). Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. *PloS one*, *9*(3), e90320.

Rosen, M. A., Salas, E., Wilson, K. A., King, H. B., Salisbury, M., Augenstein, J. S., ... & Birnbach, D. J. (2008). Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simulation in Healthcare*, *3*(1), 33-41.

*Rothermich, A. E., & Saunders, J. M. (1977). *Team Effectiveness Rating Scale*. Unpublished instrument.

Rousseau, C., Laurin-Lamothe, A., Nadeau, L., Deshaies, S., & Measham, T. (2012). Measuring the quality of interprofessional collaboration in child mental health collaborative care. *International Journal of Integrated Care*, *12*, 1-8.

*Schroder, C., Medves, J., Paterson, M., Byrnes, V., Chapman, C., O'Riordan, A., ... & Kelly, C. (2011). Development and pilot testing of the collaborative practice assessment tool. *Journal of Interprofessional Care, 25*(3), 189-195.

*Shortell, S. M., Rousseau, D. M., Gillies, R. R., Devers, K. J., & Simons, T. L. (1991). Organizational assessment in intensive care units (ICUs): Construct development, reliability, and validity of the ICU nurse-physician questionnaire. *Medical Care 29*(8), 709-726.

Shortell, S. M., Zimmerman, J. E., Rousseau, D. M., Gillies, R. R., Wagner, D. P., Draper, E. A., ... & Duffy, J. (1994). The performance of intensive care units: does good management make a difference?. *Medical care*, 508-525.

*Singleton, A., Smith, F., Harris, T., Ross-Harper, R., & Hilton, S. (1999). An evaluation of the team objective structured clinical examination (TOSCE). *Medical Education*, *33*(1), 34-41.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300-308.

Solomon, P., Marshall, D., Boyle, A., Burns, S., Casimiro, L. M., Hall, P., & Weaver, L. (2011). Establishing face and content validity of the McMaster-Ottawa team observed structured clinical encounter (TOSCE). *Journal of Interprofessional Care*, *25*(4), 302-304.

*Taylor, C., Atkins, L., Richardson, A., Tarrant, R., & Ramirez, A. J. (2012). Measuring the quality of MDT working: an observational approach. *BMC Cancer, 12*(1), 202.

*Taylor, C., Brown, K., Lamb, B., Harris, J., Sevdalis, N., & Green, J. S. A. (2012). Developing and testing TEAM (Team Evaluation and Assessment Measure), a self-assessment tool to improve cancer multidisciplinary teamwork. *Annals of Surgical Oncology, 19*(13), 4019-4027.

*Temkin-Greener, H., Gross, D., Kunitz, S. J., & Mukamel, D. (2004). Measuring interdisciplinary team performance in a long-term care setting. *Medical Care, 42*(5), 472-481.

*Thompson, B. M., Levine, R. E., Kennedy, F., Naik, A. D., Foldes, C. A., Coverdale, J. H., ... & Haidet, P. (2009). Evaluating the quality of learning-team processes in medical education: development and validation of a new measure. *Academic Medicine, 84*(10), S124-S127.

Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base.* Cincinnati, OH: Atomic Dog.

*Tsukuda, R. A., & Stahelski, A. J. (1990). *Team Skills Questionnaire*. Unpublished instrument.

*Upenieks, V. V., Lee, E. A., Flanagan, M. E., & Doebbeling, B. N. (2010). Healthcare Team Vitality Instrument (HTVI): Developing a tool assessing healthcare team functioning. *Journal of Advanced Nursing, 66*(1), 168-176.

*Varney, G. H. (1991). Building productive teams: An action guide and resource book (pp. 31-32). San Francisco, CA: Jossey-Bass.

*Walker, S., Brett, S., McKay, A., Lambden, S., Vincent, C., & Sevdalis, N. (2011). Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR): Development and validation. *Resuscitation*, *82*(7), 835-844.

*Wallin, C. J., Meurling, L., Hedman, L., Hedegård, J., & Felländer-Tsai, L. (2007). Target-focused medical emergency team training using a human patient simulator: Effects on behaviour and attitude. *Medical Education, 41*(2), 173-180.

Weaver, S. J., Lyons, R., DiazGranados, D., Rosen, M. A., Salas, E., Oglesby, J., ... & King, H. B. (2010). The anatomy of health care team training and the state of practice: a critical review. *Academic Medicine*, *85*(11), 1746-1760.

*Weller, J., Frengley, R., Torrie, J., Shulruf, B., Jolly, B., Hopley, L., ... & Paul, A. (2011). Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. *BMJ Quality & Safety*, *20,* 216-222.

Weller, J., Shulruf, B., Torrie, J., Frengley, R., Boyd, M., Paul, A., ... & Dzendrowskyj, P. (2013). Validation of a measurement tool for self-assessment of teamwork in intensive care. *British Journal of Anaesthesia*, 1-8.

Wheelan, S. A., & Hochberger, J. M. (1993). *The Group Development Questionnaire.* Philadelphia: GDQ Associates.

Wheelan, S. A., & Hochberger, J. M. (1996). Validation studies of the group development questionnaire. *Small Group Research, 27*(1), 143-170.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World Journal of Surgery*, *32*(4), 548-556.

*Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Medical Education, 40*(11), 1098-1104.

Appendix A

Health Care Team Performance Measures and General Characteristics

Reference	Measure Name	Accessibility	Clarity of Language	Instrument Type	Applicability	Objectivity vs. Bias
Agency for Healthcare Research and Quality (AHRQ; 2012)	Team Assessment Questionnaire (TAQ)	Open access: http://www.ahrq.gov/sites/default/files/wysiwyg/professionals/education/curriculum-tools/teamstepps/instructor/reference/tmassess.pdf	High	Self-report	Adaptable	Less Objective
AHRQ (2014)	Teamwork Perceptions Questionnaire (T-TPQ)	Open access: http://www.ahrq.gov/professionals/education/curriculum-tools/teamstepps/instructor/reference/teampercept.html	High	Self-report	Adaptable	Less Objective
AHRQ (2014)	Team Performance Observation Tool (TPOT)	Open access: http://www.ahrq.gov/sites/default/files/wysiwyg/professionals/education/curriculum-tools/teamstepps/instructor/reference/tmpot.pdf	Moderate	Observer	Adaptable	Very Objective
Anderson & West (1998)	Team Climate Inventory	Copyrighted	High	Self-report	Generic	Less Objective
Archibald et al. (2014)	The Interprofessional Collaborative Competency Attainment Survey (ICCAS)	Subscription required	High	Self-report	Adaptable	Less Objective

Baggs (1994)	Collaboration and Satisfaction about Care Decisions	Subscription required	High	Self-report	Generic	Less objective
Bailey et al. (1983)	Rating Individual Participation in Teams	Unpublished	High	Observer	Generic	Very objective
Batorowicz & Shepherd (2008)	Team Decision Making Questionnaire (TMDQ)	Subscription required	High	Self-report	Adaptable	Less objective
Cooper et al. (2010)	Team Emergency Assessment Measure (TEAM)	Open access: http://www.midss.org/sites/default/files/final_team_tool_0.pdf	High	Observer	Generic	Fairly objective
Dyer (1987)	Team Development Scale	Copyrighted	High	Self-report	Generic	Less objective
Farrell et al. (2001)	Team Anomie Scale	Unpublished	High	Self-report	Generic	Less objective
Farrell et al. (2008)	Family Medicine Medication Use Processes Matrix (MUPM)	Subscription required	High	Observer	Adaptable	Fairly objective
Fletcher et al. (2003)	Anesthetists' Non-technical Skills (ANTS) behavioral marker system	Copyrighted	Moderate	Observer	Generic	Fairly objective
Healey et al. (2004)	The Observational Teamwork Assessment for Surgery (OTAS)	Copyrighted	High	Observer	Focused	Very objective
Heinemann et al. (1999)	Attitude Toward Health Care Teams	Copyrighted	High	Self-report	Adaptable	Less objective
Hepburn et al. (1998)	Team Skills Scale	Copyrighted	High	Self-report	Focused	Less objective
Hojat et al. (1999)	Jefferson Scale of Attitudes Toward Nurse-Physician Collaboration	Subscription required	High	Self-report	Adaptable	Less objective
Kenaszchuk et al. (2010)	Adapted version of Nurses'	Open access: http://www.biomedc	High	Self-report	Adaptable	Less objective

	Opinion Questionnaire (NOQ) of the Ward Organisa-tional Features Scales	entral.com/1472-6963/10/83				
Kiesewetter & Fischer (2015)	The Teamwork Assessment Scale (TAS)	Open access: http://www.egms.de/static/en/journals/zma/2015-32/zma000961.shtml	High	Observer	Adaptable	Fairly objective
King et al. (2010)	The Interprofessional Socialization and Valuing Scale (ISVS)	Subscription required	High	Self-report	Generic	Less objective
Lamb et al. (2011)	Multidisciplinary Team Performance Tool	Subscription required	High	Observer	Focused	Fairly objective
Lazar (1971)	Team Excellence Questionnaire	Copyrighted	High	Self-report	Generic	Less objective
Lazar (1985)	Factors Influencing Productivity and Excellence of Team Work	Copyrighted	High	Self-report	Generic	Less objective
Lichtenstein et al. (1997)	Team Integration Measure	Subscription required	High	Self-report	Generic	Less objective
Lyk-Jensen et al. (2014)	Nurse Anesthetists' Non-Technical Skills (N-ANTS)	Copyrighted	High	Observer	Focused	Fairly objective
Malec et al. (2007)	Mayo High Performance Teamwork Scale (MHPTS)	Subscription required	High	Observer	Adaptable	Less objective
McClane (1992)	Team Assessment Worksheets	Copyrighted	High	Self-report	Generic	Less objective
Mishra et al. (2009)	The Oxford Non-Technical Skills (NOTECHS)	Subscription required	High	Observer	Adaptable	Fairly objective

Moos (1986)	Group Environment Scale	Copyrighted	High	Self-report	Generic	Less objective
National Patient Safety Agent (2006)	The Team Climate Assessment Measurement (TCAM)	Open access: http://www.nrls.npsa.nhs.uk/resources/?entryid45=59884	High	Self-report	Generic	Less objective
Norris et al. (2015)	Interprofessional Attitudes Scale (IPAS)	Open access: https://nexusipe-resource-exchange.s3.amazonaws.com/Interprofessional%20Attitudes%20Scale%20(IPAS)_0.pdf	High	Self-report	Adaptable	Less objective
Ødegård (2006)	Perception of Interprofessional Collaboration Questionnaire (PINCOM-Q)	Subscription required	High	Self-report	Generic	Less objective
Orchard et al. (2012)	Assessment of Interprofessional Team Collaboration Scale (AITCS)	Subscription required	High	Self-report	Adaptable	Less objective
Ottestad et al. (2007)	Unnamed scale	Subscription required	Low	Observer	Focused	Very objective
Parsell & Bligh (1999)	Readiness of Health Care Students for Interprofessional Learning (RIPLS)	Subscription required	High	Self-report	Focused	Less objective
Pollard, Miers, & Gilchrist (2004)	UWE Entry Level Interprofessional Questionnaire, ELIQ	Subscription required	High	Self-report	Adaptable	Less objective
Robertson et al. (2014)	The Oxford Non-Technical Skills (NOTECHS) II	Open access: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090320	High	Observer	Adaptable	Fairly objective
Rothermich & Saunders (1977)	Team Effectiveness Rating Scale	Unpublished	High	Self-report	Generic	Less objective

Schroder et al. (2011)	Collaborative Practice Assessment Tool (CPAT)	Subscription required	High	Self-report	Adaptable	Less objective
Shortell et al. (1991)	Intensive Care Unit Nurse/Physician Instrument	Subscription required	High	Self-report	Adaptable	Less objective
Singleton et al. (1999)	McMaster-Ottawa Team Observed Structured Clinical Encounter (TOSCE)	Open access: http://fhs.mcmaster.ca/tosce/en/	Moderate	Observer	Generic	Fairly objective
Taylor, Atkins et al. (2012)	Multidisciplinary team observational assessment rating scale (MDT-OARS)	Subscription required	High	Observer	Adaptable	Very objective
Taylor, Brown et al. (2012)	Team Evaluation and Assessment Measure (TEAM)	Subscription required	High	Self-report	Focused	Less objective
Temkin-Greener et al. (2004)	Interdisciplinary Team Performance Scale (ITPS)	Subscription required	High	Self-report	Adaptable	Less objective
Thompson et al. (2009)	Team Performance Scale (TPS)	Subscription required	High	Self-report	Generic	Less objective
Tsukuda & Stahelski (1990)	Team Skills Questionnaire	Copyrighted	High	Self-report	Generic	Less objective
Upenieks et al. (2010)	Healthcare Team Vitality Instrument	Open access: http://www.ihi.org/resources/Pages/Tools/HealthcareTeamVitalityInstrument.aspx	High	Self-report	Adaptable	Less objective
Varney (1991)	Analyzing Team Effectiveness	Copyrighted	High	Self-report	Generic	Less objective
Walker et al. (2011)	Observational Skill-based Clinical Assess-ment tool for Resuscitation (OSCAR)	Subscription	Moderate	Observer	Focused	Fairly objective
Wallin et al. (2007)	Emergency medicine crisis	Subscription required	High	Observer	Generic	Fairly objective

	resource management (EMCRM)					
Weller et al. (2011)	Modified Version of the Mayo High Performance Teamwork Scale	Subscription required	High	Observer	Adaptable	Fairly objective
Wheelan & Hochberger (1996)	Group Development Questionnaire (GDQ)	Copyrighted	High	Self-report	Generic	Less objective
Yule et al. (2006)	Non-technical Skills for Surgeons (NOTTs) Rating Scale	Copyrighted	High	Observer	Focused	Very objective

Appendix B

Health Care Team Performance Measures and Reliability and Validity Information

Reference	Measure Name	Criterion Validity	Construct Validity	Content Validity	Inter-Rater or Inter-Observer Reliability	Internal Consistency	Test-Retest Reliability	Additional Related Citations
Agency for Healthcare Research and Quality (AHRQ; 2012)	Team Assessment Questionnaire (TAQ)	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive literature review and pilot testing	N/A (self-report)	High: total scale Cronbach's α was .93	Unable to find supporting evidence	Beebe et al. (2012)
AHRQ (2014)	Teamwork Perceptions Questionnaire (T-TPQ)	Unable to find supporting evidence	Moderate evidence: confirmatory factor analysis (CFA) conducted and supported theorized structure	Detailed information available: items based on extensive literature review	N/A (self-report)	High: Cronbach's α ranged from .92-.96 for subscales	Unable to find supporting evidence	Keebler et al. (2014)
AHRQ (2014)	The Trauma Team Performance Observation Tool (TPOT)	Moderate evidence: concurrent validity evidence (e.g., negatively correlated with number of medical errors)	Moderate evidence: significant difference in pre and post training scores following team training	Detailed information available: items based on extensive literature review; interviews, expert review, and observed behavior	Low: the average intraclass correlation (ICC) was .54 and the average level of agreement was 75%	Low: Cronbach's α ranged from .53 to .64 for subscales	High: kappa = .71	Beebe et al. (2012); Capella et al. (2010)
Anderson & West (1998)	Team Climate Inventory	Moderate evidence:	Strong evidence: a series of one-	Detailed information	N/A (self-report)	High: Cronbach's	Unable to find	Heinemann & Zeiss (2002)

		predictive validity evidence (e.g., predicted number of innovations)	way analysis of variances (ANOVAs) conducted and indicate significant differences among five samples of teams in expected manner	available: items based on extensive literature review and previous measures		α ranged from .84-.94 for subscales	supporting evidence	
Archibald et al. (2014)	The Interprofessional Collaborative Competency Attainment Survey (ICCAS)	Unable to find supporting evidence	Strong evidence: exploratory factor analysis (EFA) conducted and significant difference in pre and post training scores following training	Detailed information available: items based on expert review and previous measures	N/A (self-report)	High: Cronbach's α ranged from .94-.96 for subscales	Unable to find supporting evidence	
Baggs (1994)	Collaboration and Satisfaction about Care Decisions	Moderate evidence: concurrent validity evidence (e.g., correlated with global collaboration score)	Moderate evidence: convergent validity evidence (i.e., correlated with satisfaction scale)	Detailed information available: items based on extensive literature review, expert review, and pilot testing	N/A (self-report)	High: total scale Cronbach's α was .93	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Bailey et al. (1983)	Rating Individual Participation in Teams	Moderate evidence: concurrent validity evidence (correlated with measure of participation)	Moderate evidence: one-way ANOVA conducted and indicated significant differences among teams in expected manner	Some information available: items based on expert review	Moderate: level of agreement was 64%	Unable to find supporting evidence	Low: developers found considerable variability across time	Heinemann & Zeiss (2002)
Batorowicz &	Team Decision Making	Unable to find	Moderate	Some	N/A (self-	High: total	Low: ICCs	

Shepherd (2008)	Questionnaire (TMDQ)	supporting evidence	evidence: principal component analysis (PCA) conducted and supported theorized structure	information available: items based on focus group interviews	report)	scale Cronbach's α was .96	ranged from .52-.94	
Cooper et al. (2010)	Team Emergency Assessment Measure	Moderate evidence: concurrent validity evidence (item to global ratings correlated strongly from videoed events)	Strong evidence: PCA conducted and additional PCA conducted with additional sample, supported theorized structure	Detailed information available: items based on extensive review of the literature, expert review, and content validity index (CVI) calculated (all items greater than .83)	Moderate: mean ICC was .6; Kappa was .55	High: total scale Cronbach's α was .97 (hospital events); .98 (simulated events)	Moderate: kappa was .53	Cooper & Cant (2014)
Dyer (1987)	Team Development Scale	Unable to find supporting evidence	Moderate evidence: size of team influenced score, as expected	Some information available: items based one expert review	N/A (self-report)	Low: Cronbach's α ranged from .47-.90 for the subscales	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Farrell et al. (2001)	Team Anomie Scale	Unable to find supporting evidence	Moderate evidence: convergent validity evidence (e.g., correlated strongly with cohesion scale)	Detailed information available: items based on extensive review of the literature, interviews, pilot testing, and observed	N/A (self-report)	High: total scale Cronbach's α was .90	Unable to find supporting evidence	Heinemann & Zeiss (2002)

				behavior				
Farrell et al. (2008)	Family Medicine Medication Use Processes Matrix (MUPM)	Unable to find supporting evidence	Moderate evidence: one-way ANOVA conducted and indicated significant differences among teams in expected manner	Detailed information available: items based on extensive literature review and expert review	Unable to find supporting evidence	High: total scale Cronbach's α was .97	Moderate: test-retest ICCs ranged from .65 to .97	Farrell et al. (2010)
Fletcher et al. (2003)	Anesthetists' Non-technical Skills (ANTS) behavioral marker system	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive literature review, interviews, pilot testing, and observed behavior	Low: r_{wg} ranged from .55 to .67	High: Cronbach's α ranged from .79 to .86 for subscales	Unable to find supporting evidence	
Healey et al. (2004)	The Observational Teamwork Assessment for Surgery (OTAS)	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive literature review and observed behavior	Unable to find supporting evidence	Unable to find supporting evidence	Unable to find supporting evidence	
Heinemann et al. (1999)	Attitude Toward Health Care Teams	Moderate evidence: concurrent validity evidence (e.g., correlated with another attitudes toward health care scale)	Strong evidence: PCA conducted and ANOVAs conducted and indicate significant differences among teams in expected manner	Detailed information available: items based on expert review and content validity index (CVI) calculated	N/A (self-report)	High: Cronbach's α ranged from .75 to .83	Low: test-retest correlation ranged from .36 to .71 for subscales	Heinemann & Zeiss (2002)

				(.95)				
Hepburn et al. (1998)	Team Skills Scale	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on expert review and previous measure	N/A (self-report)	High: total scale Cronbach's α was .94	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Hojat et al. (1999)	Jefferson Scale of Attitudes Toward Nurse-Physician Collaboration	Unable to find supporting evidence	Strong evidence: PCA conducted and supported theorized structure and ANOVAs conducted and indicate significant differences among teams in expected manner	Some information available: items based on a previous measure	N/A (self-report)	High: total scale Cronbach's α was .84 with one sample (medical students), .85 with an additional sample (nursing students)	Unable to find supporting evidence	Hojat et al. (2001); Hojat et al. (2003)
Kenaszchuk et al. (2010)	Adapted version of Nurses' Opinion Questionnaire (NOQ) of the Ward Organisa-tional Features Scales	Moderate evidence: concurrent validity evidence (e.g., performed pairwise hospital site comparisons of mean scale score for NWI-NPRS among ratings)	Strong evidence: CFA conducted, convergent validity evidence (e.g., correlated with the Collegial Nurse-Physician Relations Subscale of the Nursing Work Index), and discriminant validity evidence (e.g., correlated with the Attitudes Toward Health Care Teams Scale)	Detailed information available: items based on extensive literature review and previous measure	N/A (self-report)	High: Cronbach's α ranged from .71 to .88 for subscales	Unable to find supporting evidence	

Kiesewetter & Fischer (2015)	The Teamwork Assessment Scale (TAS)	Moderate evidence: concurrent validity evidence (e.g., correlated with clinical performance)	Moderate evidence: EFA conducted and supported theorized structure but could not differentiate between expected two dimensions	Detailed information available: items based on extensive literature review, expert review, and pilot testing	Unable to find supporting evidence	Moderate: Cronbach's α ranged from .67 to .81 for subscales	Unable to find supporting evidence	
King et al. (2010)	The Interprofessional Socialization and Valuing Scale (ISVS)	Unable to find supporting evidence	Moderate evidence: PCA conducted and supported theorized structure	Detailed information available: items based on extensive literature review and expert review	N/A (self-report)	High: Cronbach's α ranged from .79 to .89 for subscales	Unable to find supporting evidence	
Lamb et al. (2011)	Multidisciplinary Team Performance Tool	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive review of the literature, expert review, and previous measure	Low: ICCs ranged from .31 to .87	Unable to find supporting evidence	Unable to find supporting evidence	
Lazar (1971)	Team Excellence Questionnaire	Unable to find supporting evidence	Unable to find supporting evidence	Unable to find supporting evidence	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Lazar (1985)	Factors Influencing Productivity and Excellence of Team Work	Unable to find supporting evidence	Unable to find supporting evidence	Some information available: items based on expert review	N/A (self-report)	Unable to find supporting evidence	High: a Wilcoxon Rank Sum Test ($p > .58$) indicated no significant	Heinemann & Zeiss (2002)

							differences between sets of scores taken at different times	
Lichtenstein et al. (1997)	Team Integration Measure	Unable to find supporting evidence	Strong evidence: correlations supported theorized relationships and discriminant validity evidence (e.g., negatively correlated with age)	Some information available: items based on previous scales	N/A (self-report)	High: Cronbach's α ranged from .90 to .91 for subscales	Unable to find supporting evidence	
Lyk-Jensen et al. (2014)	Nurse Anesthetists' Non-Technical Skills (N-ANTS)	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive review of the literature, interviews and expert review	Unable to find supporting evidence	Unable to find supporting evidence	Unable to find supporting evidence	
Malec et al. (2007)	Mayo High Performance Teamwork Scale (MHPTS)	Unable to find supporting evidence	Strong evidence: significant difference in pre and post training scores following team training and additional evidence via Rasch indicators	Some information available: based on extensive review of the literature	High: Rasch Person reliability ranged from .71 to .79	High: Cronbach's α was .85 for all ratings	Unable to find supporting evidence	
McClane (1992)	Team Assessment Worksheets	Unable to find supporting evidence	Unable to find supporting evidence	Unable to find supporting evidence	N/A (self-report)	High: Cronbach's α ranged from .85 to .95 for	Unable to find supporting evidence	Heinemann & Zeiss (2002)

						subscales		
Mishra et al. (2009)	The Oxford Non-Technical Skills (NOTECHS)	Moderate: concurrent validity evidence (e.g., correlated with technical error)	Strong evidence: significant difference in pre and post training scores following training (multiple studies)	Detailed information available: based on extensive review of the literature, expert review, and previous measure	High: R_{wg} was .99	Unable to find supporting evidence	High: an ANOVA indicated no significant difference between sets of scores taken at different times	
Moos (1986)	Group Environment Scale	Strong: concurrent (e.g., correlated with cohesion; Evan & Jarris, 1986) and predictive validity (e.g., predicted organizational functioning; Giamartino, 1981) evidence	Strong evidence: scale indicates significant differences among teams in expected manner (multiple studies)	Detailed information available: based on interviews and observed behaviors	N/A (self-report)	Moderate: Cronbach's α ranged from .62 to .86 for subscales	Moderate: ranges from .65 to .87	Heinemann & Zeiss (2002); Littlepage et al. (1989); Moos (1994)
National Patient Safety Agent (2006)	The Team Climate Assessment Measurement (TCAM)	Unable to find supporting evidence	Unable to find supporting evidence	Unable to find supporting evidence	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	
Norris et al. (2015)	Interprofessional Attitudes Scale (IPAS)	Unable to find supporting evidence	Strong evidence: CFA conducted and supported theorized structure and additional EFA conducted (responses randomly split)	Detailed information available: items based on expert review and previous measure	N/A (self-report)	Moderate: Cronbach's α ranged from .62 to .92 for subscales	Unable to find supporting evidence	
Ødegård (2006)	Perception of Interprofessional	Moderate evidence:	Strong evidence: PCA conducted	Detailed information	N/A (self-report)	Low: Cronbach's	Unable to find	Ødegård et al. (2008);

	Collaboration Questionnaire(PINCOM-Q)	concurrent validity evidence (e.g., correlated with EDC-P)	and supported theorized structure and g-test completed	available: items based on expert review and previous measure		α ranged from .55 to .82 for subscales	supporting evidence	Rousseau et al. (2012)
Orchard et al. (2012)	Assessment of Interprofessional Team Collaboration Scale (AITCS)	Unable to find supporting evidence	Moderate evidence: PCA conducted and supported theorized structure	Detailed information available: items based on extensive literature review and expert review	N/A (self-report)	High: Cronbach's α ranged from .80 to .97 for subscales	Unable to find supporting evidence	
Ottestad et al. (2007)	Unnamed scale	Moderate evidence: concurrent validity evidence (e.g., correlated with nontechnical scores)	Unable to find supporting evidence	Unable to find supporting evidence	High: interrater reliability was .88	Unable to find supporting evidence	Unable to find supporting evidence	
Parsell & Bligh (1999)	Readiness of Health Care Students for Interprofessional Learning (RIPLS)	Unable to find supporting evidence	Strong: PCA conducted and supported theorized structure (multiple studies)	Detailed information available: items based on extensive review of the literature and expert review	N/A (self-report)	High: total scale Cronbach's α was .90	Unable to find supporting evidence	Reid et al. (2006)
Pollard, Miers, & Gilchrist (2004)	UWE Entry Level Interprofessional Questionnaire, ELIQ	Moderate evidence: concurrent validity evidence (e.g., correlated with IEPS; Leucht et al., 1990)	Strong evidence: EFA conducted and supported theorized structure and significant differences between samples in expected manners	Detailed information available: items based on extensive review of the literature and pilot testing	N/A (self-report)	High: total scale Cronbach's α was .71	High: test-retest reliability ranged from .77 to .86	Pollard et al. (2005)

Robertson et al. (2014)	The Oxford Non-Technical Skills (NOTECHS) II	Moderate evidence: concurrent validity evidence (e.g., correlated with WHO time-out)	Unable to find supporting evidence	Detailed information available: items based on expert review, previous measure, and pilot testing	Low: levels of agreement ranged between 45% and 78%	Unable to find supporting evidence	Unable to find supporting evidence	Morgan et al. (2015)
Rothermich & Saunders (1977)	Team Effectiveness Rating Scale	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive review of the literature, expert review, and previous measures	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Schroder et al. (2011)	Collaborative Practice Assessment Tool (CPAT)	Unable to find supporting evidence	Strong evidence: EFA conducted and CFA conducted and supported theorized structure (multiple studies)	Detailed information available: items based on extensive review of the literature, expert review, and pilot testing	N/A (self-report)	High: Cronbach's α ranged from .73 to .84 for subscales	Unable to find supporting evidence	
Shortell et al. (1991)	Intensive Care Unit Nurse/Physician Instrument	Moderate evidence: concurrent validity evidence (e.g., correlated with coordination)	Strong evidence: convergent validity evidence (e.g., correlated with communication) and discriminant validity evidence (e.g., negatively correlated with	Detailed information available: items based on pilot testing	N/A (self-report)	Moderate: Cronbach's α ranged from .64 to .94 for subscales	Unable to find supporting evidence	Heinemann & Zeiss (2002); Shortell et al. (1994)

			turnover)					
Singleton et al. (1999)	McMaster-Ottawa Team Observed Structured Clinical Encounter (TOSCE)	Unable to find supporting evidence	Moderate evidence: generalizability study (G-study) completed	Detailed information available: items based on expert review and observed behavior	High: small variance in ratings	Unable to find supporting evidence	Unable to find supporting evidence	Lie et al. (2015); Solomon et al. (2011)
Taylor, Atkins et al. (2012)	Multidisciplinary team observational assessment rating scale (MDT-OARS)	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on extensive literature review, observed behavior, and expert review	Low: ICCs ranged from .32-.92	Unable to find supporting evidence	Unable to find supporting evidence	
Taylor, Brown et al. (2012)	Team Evaluation and Assessment Measure (TEAM)	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on expert review and pilot testing	N/A (self-report)	Low: Cronbach's α ranged from .52 to .81 for subscales	Unable to find supporting evidence	
Temkin-Greener et al. (2004)	Interdisciplinary Team Performance Scale (ITPS)	Unable to find supporting evidence	Moderate evidence: regression analyses conducted supported theorized relationships	Some information available: items based on expert review	N/A (self-report)	High: Cronbach's α ranged from .76 to .89 for subscales	Unable to find supporting evidence	
Thompson et al. (2009)	Team Performance Scale (TPS)	Unable to find supporting evidence	Strong evidence: EFA conducted and ANOVAs conducted and indicate significant	Detailed information available: items based on extensive literature and	N/A (self-report)	High: total scale Cronbach's α was .97	Unable to find supporting evidence	

			differences among teams in expected manner	expert review				
Tsukuda & Stahelski (1990)	Team Skills Questionnaire	Unable to find supporting evidence	Unable to find supporting evidence	Some information available: based on expert review	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Upenieks et al. (2010)	Healthcare Team Vitality Instrument (HTVI)	Unable to find supporting evidence	Strong evidence: convergent validity evidence and CFA conducted	Detailed information available: items based on extensive literature review, expert review, and previous measures	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	
Varney (1991)	Analyzing Team Effectiveness	Unable to find supporting evidence	Unable to find supporting evidence	Some information available: items based on extensive literature review	N/A (self-report)	Unable to find supporting evidence	Unable to find supporting evidence	Heinemann & Zeiss (2002)
Walker et al. (2011)	Observational Skill-based Clinical Assess-ment tool for Resuscitation (OSCAR)	Moderate evidence: concurrent validity evidence (e.g., correlated with TEAM measure)	Unable to find supporting evidence	Detailed information available: items based on extensive literature review, expert review, and previous measures	Moderate: ICCs ranged from .61 to .88	High: Cronbach's α ranged from .74 to .97 for subscales	Unable to find supporting evidence	McKay et al. (2012)
Wallin et al. (2007)	Emergency medicine crisis resource management (EMCRM)	Unable to find supporting evidence	Moderate evidence: significant	Detailed information available:	Moderate: inter-rater reliability	Unable to find supporting	Unable to find supporting	

			difference in pre and post training scores following team training	items based on observed behavior and expert review	ranged from .60 to .78	evidence	evidence	
Weller et al. (2011)	Modified Version of the Mayo High Performance Teamwork Scale	Unable to find supporting evidence	Strong evidence: EFA conducted and significant difference in scores over time	Detailed information available: items based on expert review and previous measure	Unable to find supporting evidence	High: Cronbach's α ranged from .89 to .92 for subscales	Unable to find supporting evidence	Weller et al. (2013)
Wheelan & Hochberger (1993)	Group Development Questionnaire (GDQ)	Moderate evidence: concurrent validity evidence (e.g., correlated with Group Attitude Scale; Evans & Jarvis, 1986)	Strong evidence: scale indicates significant differences among teams in expected manner (multiple studies)	Detailed information available: items based on extensive literature review, expert review, and previous measures	N/A (self-report)	Moderate: Cronbach's α ranged from .69 to .88 for subscales	Moderate: test-retest reliability ranged from .69 to .82 for subscales	Heinemann & Zeiss (2002); Wheelan & Hochberger (1996)
Yule et al. (2006)	Non-technical Skills for Surgeons (NOTTs) Rating Scale	Unable to find supporting evidence	Unable to find supporting evidence	Detailed information available: items based on interviews, observed behavior, and expert review	Low: mean r_{wg} ranged from .46-.74	High: mean absolute difference between raters' element ratings and categories indicated high consistency	Unable to find supporting evidence	Yule et al. (2008)

NOTE: Some measures excluded because we could not access ANY information about them (Helmreich's ORMAQ)