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
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 measure 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 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*

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Unpublished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

<table>
<thead>
<tr>
<th>Summary of Available Validity Information related to Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>Criterion Validity</td>
</tr>
<tr>
<td>High evidence</td>
</tr>
<tr>
<td>Some evidence</td>
</tr>
<tr>
<td>No evidence identified</td>
</tr>
<tr>
<td>Construct Validity</td>
</tr>
<tr>
<td>High evidence</td>
</tr>
<tr>
<td>Some evidence</td>
</tr>
<tr>
<td>No evidence identified</td>
</tr>
<tr>
<td>Content Validity</td>
</tr>
<tr>
<td>Detailed information available</td>
</tr>
<tr>
<td>Some information available</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Characteristic</th>
<th>Number of Articles (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inter-Rater Reliability</strong></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4 (21.1%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>4 (21.1%)</td>
</tr>
<tr>
<td>Low</td>
<td>6 (31.6%)</td>
</tr>
<tr>
<td>No information identified</td>
<td>5 (26.3%)</td>
</tr>
<tr>
<td>N/A</td>
<td>34 (64.2%)</td>
</tr>
<tr>
<td><strong>Internal Consistency</strong></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>27 (50.9%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>5 (9.4%)</td>
</tr>
<tr>
<td>Low</td>
<td>4 (7.5%)</td>
</tr>
<tr>
<td>No information identified</td>
<td>17 (32.1%)</td>
</tr>
<tr>
<td><strong>Test-retest Reliability</strong></td>
<td></td>
</tr>
</tbody>
</table>
**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.

<table>
<thead>
<tr>
<th>Level</th>
<th>Count (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>4 (7.5%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>4 (7.5%)</td>
</tr>
<tr>
<td>Low</td>
<td>3 (5.7%)</td>
</tr>
<tr>
<td>No information identified</td>
<td>42 (79.2%)</td>
</tr>
</tbody>
</table>
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
  o 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


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


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


### Appendix A

#### Health Care Team Performance Measures and General Characteristics

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure Name</th>
<th>Accessibility</th>
<th>Clarity of Language</th>
<th>Instrument Type</th>
<th>Applicability</th>
<th>Objectivity vs. Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson &amp; West (1998)</td>
<td>Team Climate Inventory</td>
<td>Copyrighted</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less Objective</td>
</tr>
<tr>
<td>Archibald et al. (2014)</td>
<td>The Interprofessional Collaborative Competency Attainment Survey (ICCAS)</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less Objective</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Title</td>
<td>Access</td>
<td>Level</td>
<td>Mode</td>
<td>Adaptability</td>
<td>Objectivity</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
<td>-------</td>
<td>------------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Baggs (1994)</td>
<td>Collaboration and Satisfaction about Care Decisions</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Cooper et al. (2010)</td>
<td>Team Emergency Assessment Measure (TEAM)</td>
<td>Open access: <a href="http://www.midss.org/sites/default/files/final_team_tool_0.pdf">http://www.midss.org/sites/default/files/final_team_tool_0.pdf</a></td>
<td>High</td>
<td>Observer</td>
<td>Generic</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Farrell et al. (2001)</td>
<td>Team Anomie Scale</td>
<td>Unpublished</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Fletcher et al. (2003)</td>
<td>Anesthetists’ Non-technical Skills (ANTS) behavioral marker system</td>
<td>Copyrighted</td>
<td>Moderate</td>
<td>Observer</td>
<td>Generic</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Heinemann et al. (1999)</td>
<td>Attitude Toward Health Care Teams</td>
<td>Copyrighted</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Hepburn et al. (1998)</td>
<td>Team Skills Scale</td>
<td>Copyrighted</td>
<td>High</td>
<td>Self-report</td>
<td>Focused</td>
<td>Less objective</td>
</tr>
<tr>
<td>Study</td>
<td>Measurement Instrument</td>
<td>Access Type</td>
<td>Frequency</td>
<td>Data Collector</td>
<td>Adaptability</td>
<td>Objectivity</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------------------------</td>
<td>------------------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>King et al. (2010)</td>
<td>The Interprofessional Socialization and Valuing Scale (ISVS)</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Lamb et al. (2011)</td>
<td>Multidisciplinary Team Performance Tool</td>
<td>Subscription required</td>
<td>High</td>
<td>Observer</td>
<td>Focused</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Lazar (1971)</td>
<td>Team Excellence Questionnaire</td>
<td>Copylefted</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Lichtenstein et al. (1997)</td>
<td>Team Integration Measure</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Malec et al. (2007)</td>
<td>Mayo High Performance Teamwork Scale (MHPTS)</td>
<td>Subscription required</td>
<td>High</td>
<td>Observer</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Mishra et al. (2009)</td>
<td>The Oxford Non-Technical Skills (NOTECHS)</td>
<td>Subscription required</td>
<td>High</td>
<td>Observer</td>
<td>Adaptable</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Reference</td>
<td>Scale Name</td>
<td>Accessibility</td>
<td>Self-report</td>
<td>Generic</td>
<td>Objective</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>---------------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Moos (1986)</td>
<td>Group Environment Scale</td>
<td>Copyrighted</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Orchard et al. (2012)</td>
<td>Assessment of Interprofessional Team Collaboration Scale (AITCS)</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Ottestad et al. (2007)</td>
<td>Unnamed scale</td>
<td>Subscription required</td>
<td>Low</td>
<td>Observer</td>
<td>Focused</td>
<td>Very objective</td>
</tr>
<tr>
<td>Pollard, Miers, &amp; Gilchrist (2004)</td>
<td>UWE Entry Level Interprofessional Questionnaire, ELI IQ</td>
<td>Subscription required</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Robertson et al. (2014)</td>
<td>The Oxford Non-Technical Skills (NOTECHS) II</td>
<td>Open access: <a href="http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090320">link</a></td>
<td>High</td>
<td>Observer</td>
<td>Adaptable</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Rothermich &amp; Saunders (1977)</td>
<td>Team Effectiveness Rating Scale</td>
<td>Unpublished</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
</tbody>
</table>
### HEALTH CARE TEAM PERFORMANCE

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Tool Description</th>
<th>Accessibility</th>
<th>Frequency</th>
<th>Methodology</th>
<th>Adaptability</th>
<th>Objectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schroder et al. (2011)</td>
<td>Collaborative Practice Assessment Tool (CPAT)</td>
<td>Subscription</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Shortell et al. (1991)</td>
<td>Intensive Care Unit Nurse/Physician Instrument</td>
<td>Subscription</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Singleton et al. (1999)</td>
<td>McMaster-Ottawa Team Observed Structured Clinical Encounter (TOSCE)</td>
<td>Open access</td>
<td>Moderate</td>
<td>Observer</td>
<td>Generic</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Taylor, Atkins et al. (2012)</td>
<td>Multidisciplinary team observational assessment rating scale (MDT-OARS)</td>
<td>Subscription</td>
<td>High</td>
<td>Observer</td>
<td>Adaptable</td>
<td>Very objective</td>
</tr>
<tr>
<td>Taylor, Brown et al. (2012)</td>
<td>Team Evaluation and Assessment Measure (TEAM)</td>
<td>Subscription</td>
<td>High</td>
<td>Self-report</td>
<td>Focused</td>
<td>Less objective</td>
</tr>
<tr>
<td>Temkin-Greener et al. (2004)</td>
<td>Interdisciplinary Team Performance Scale (ITPS)</td>
<td>Subscription</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Thompson et al. (2009)</td>
<td>Team Performance Scale (TPS)</td>
<td>Subscription</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Tsukuda &amp; Stahelski (1990)</td>
<td>Team Skills Questionnaire</td>
<td>Copyrighted</td>
<td>High</td>
<td>Self-report</td>
<td>Generic</td>
<td>Less objective</td>
</tr>
<tr>
<td>Upenieks et al. (2010)</td>
<td>Healthcare Team Vitality Instrument</td>
<td>Open access</td>
<td>High</td>
<td>Self-report</td>
<td>Adaptable</td>
<td>Less objective</td>
</tr>
<tr>
<td>Walker et al. (2011)</td>
<td>Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR)</td>
<td>Subscription</td>
<td>Moderate</td>
<td>Observer</td>
<td>Focused</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Wallin et al. (2007)</td>
<td>Emergency medicine crisis</td>
<td>Subscription</td>
<td>High</td>
<td>Observer</td>
<td>Generic</td>
<td>Fairly objective</td>
</tr>
<tr>
<td>Reference</td>
<td>Tool Description</td>
<td>Access Required</td>
<td>Use</td>
<td>Adaptable</td>
<td>Objectivity</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-----------------</td>
<td>-----</td>
<td>-----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Weller et al. (2011)</td>
<td>Modified Version of the Mayo High Performance Teamwork Scale</td>
<td>Subscription required</td>
<td>Observer</td>
<td>Adaptable</td>
<td>Fairly objective</td>
<td></td>
</tr>
<tr>
<td>Yule et al. (2006)</td>
<td>Non-technical Skills for Surgeons (NOTTs) Rating Scale</td>
<td>Copyrighted</td>
<td>Observer</td>
<td>Focused</td>
<td>Very objective</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix B

#### Health Care Team Performance Measures and Reliability and Validity Information

<table>
<thead>
<tr>
<th>Reference</th>
<th>Measure Name</th>
<th>Criterion Validity</th>
<th>Construct Validity</th>
<th>Content Validity</th>
<th>Inter-Rater or Inter-Observer Reliability</th>
<th>Internal Consistency</th>
<th>Test-Retest Reliability</th>
<th>Additional Related Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency for Healthcare Research and Quality (AHRQ; 2012)</td>
<td>Team Assessment Questionnaire (TAQ)</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on extensive literature review and pilot testing</td>
<td>N/A (self-report)</td>
<td>High: total scale Cronbach’s ( \alpha ) was .93</td>
<td>Unable to find supporting evidence</td>
<td>Beebe et al. (2012)</td>
</tr>
<tr>
<td>AHRQ (2014)</td>
<td>Teamwork Perceptions Questionnaire (T-TPQ)</td>
<td>Unable to find supporting evidence</td>
<td>Moderate evidence: confirmatory factor analysis (CFA) conducted and supported theorized structure</td>
<td>Detailed information available: items based on extensive literature review</td>
<td>N/A (self-report)</td>
<td>High: Cronbach’s ( \alpha ) ranged from .92-.96 for subscales</td>
<td>Unable to find supporting evidence</td>
<td>Keebler et al. (2014)</td>
</tr>
<tr>
<td>AHRQ (2014)</td>
<td>The Trauma Team Performance Observation Tool (TPOT)</td>
<td>Moderate evidence: concurrent validity evidence (e.g., negatively correlated with number of medical errors)</td>
<td>Moderate evidence: significant difference in pre and post training scores following team training</td>
<td>Detailed information available: items based on extensive literature review; interviews, expert review, and observed behavior</td>
<td>Low: the average intraclass correlation (ICC) was .54 and the average level of agreement was 75%</td>
<td>Low: Cronbach’s ( \alpha ) ranged from .53 to .64 for subscales</td>
<td>High: kappa = .71</td>
<td>Beebe et al. (2012); Capella et al. (2010)</td>
</tr>
<tr>
<td>Authors</td>
<td>Survey/Metric</td>
<td>Findings</td>
<td>Methodology</td>
<td>Supporting Evidence</td>
<td>Rating</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>---------------------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archibald et al. (2014)</td>
<td>The Interprofessional Collaborative Competency Attainment Survey (ICCAS)</td>
<td>Unable to find evidence</td>
<td>Strong evidence: exploratory factor analysis (EFA) conducted and significant difference in pre and post training scores following training</td>
<td>N/A (self-report)</td>
<td>High: Cronbach’s α ranged from .94-.96 for subscales</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baggs (1994)</td>
<td>Collaboration and Satisfaction about Care Decisions</td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with global collaboration score)</td>
<td>Moderate evidence: convergent validity evidence (i.e., correlated with satisfaction scale)</td>
<td>N/A (self-report)</td>
<td>High: total scale Cronbach’s α was .93</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bailey et al. (1983)</td>
<td>Rating Individual Participation in Teams</td>
<td>Moderate evidence: concurrent validity evidence (correlated with measure of participation)</td>
<td>Moderate evidence: one-way ANOVA conducted and indicated significant differences among teams in expected manner</td>
<td>Some information available: items based on expert review</td>
<td>Moderate: level of agreement was 64%</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batorowicz &amp;</td>
<td>Team Decision Making</td>
<td>Unable to find evidence</td>
<td>Moderate</td>
<td>N/A (self-report)</td>
<td>High: total</td>
<td>Low: ICCs</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Some</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Measure</td>
<td>Evidence</td>
<td>Information Available</td>
<td>Report</td>
<td>Scale</td>
<td>Cronbach’s α</td>
<td>Evidence</td>
<td>Supporting Evidence</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------------------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
<td>--------</td>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Shepherd (2008)</td>
<td>Questionnaire (TMDQ)</td>
<td>supporting evidence</td>
<td>principal component analysis (PCA) conducted and supported theorized structure</td>
<td>report</td>
<td>scale</td>
<td>Cronbach’s α was .96</td>
<td>Moderate: PCA conducted</td>
<td>ranged from .52-.94</td>
</tr>
<tr>
<td>Cooper et al. (2010)</td>
<td>Team Emergency Assessment Measure</td>
<td>Moderate evidence: concurrent validity evidence (item to global ratings correlated strongly from videoed events)</td>
<td>Strong evidence: PCA conducted and additional PCA conducted with additional sample, supported theorized structure</td>
<td>Moderate: mean ICC was .6; Kappa was .55</td>
<td>High: total scale Cronbach’s α was .97 (hospital events); .98 (simulated events)</td>
<td>Moderate: kappa was .53</td>
<td>Cooper &amp; Cant (2014)</td>
<td></td>
</tr>
<tr>
<td>Dyer (1987)</td>
<td>Team Development Scale</td>
<td>Unable to find supporting evidence</td>
<td>Moderate evidence: size of team influenced score, as expected</td>
<td>N/A (self-report)</td>
<td>Low: Cronbach’s α ranged from .47-.90 for the subscales</td>
<td>Unable to find supporting evidence</td>
<td>Heinemann &amp; Zeiss (2002)</td>
<td></td>
</tr>
<tr>
<td>Farrell et al. (2001)</td>
<td>Team Anomie Scale</td>
<td>Unable to find supporting evidence</td>
<td>Moderate evidence: convergent validity evidence (e.g., correlated strongly with cohesion scale)</td>
<td>N/A (self-report)</td>
<td>High: total scale Cronbach’s α was .90</td>
<td>Unable to find supporting evidence</td>
<td>Heinemann &amp; Zeiss (2002)</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Measure/Scale</td>
<td>Supporting Evidence</td>
<td>Detailed Information</td>
<td>Comparison Evidence</td>
<td>Cronbach’s α</td>
<td>Additional Evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farrell et al. (2008)</td>
<td>Family Medicine Medication Use Processes Matrix (MUPM)</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>High: total scale Cronbach’s α was .97</td>
<td>Moderate: test-retest ICCs ranged from .65 to .97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fletcher et al. (2003)</td>
<td>Anesthetists’ Non-technical Skills (ANTS) behavioral marker system</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Low: $r_{wg}$ ranged from .55 to .67</td>
<td>High: Cronbach’s α ranged from .79 to .86 for subscales</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healey et al. (2004)</td>
<td>The Observational Teamwork Assessment for Surgery (OTAS)</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heinemann et al. (1999)</td>
<td>Attitude Toward Health Care Teams</td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with another attitudes toward health care scale)</td>
<td>Strong evidence: PCA conducted and ANOVAs conducted and indicate significant differences among teams in expected manner</td>
<td>Detailed information available: items based on expert review and content validity index (CVI) calculated</td>
<td>N/A (self-report)</td>
<td>High: Cronbach’s α ranged from .75 to .83</td>
<td>Low: test-retest correlation ranged from .36 to .71 for subscales</td>
<td>Heinemann &amp; Zeiss (2002)</td>
</tr>
<tr>
<td>Study</td>
<td>Scale/Questionnaire</td>
<td>Evidence Level</td>
<td>Supporting Evidence</td>
<td>Cronbach’s ( \alpha )</td>
<td>Type of Evidence</td>
<td>Supporting Evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepburn et al. (1998)</td>
<td>Team Skills Scale</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>( .95 )</td>
<td>N/A (self-report)</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hojat et al. (1999)</td>
<td>Jefferson Scale of Attitudes Toward Nurse-Physician Collaboration</td>
<td>Unable to find supporting evidence</td>
<td>Strong evidence: PCA conducted and supported theorized structure and ANOVAs conducted and indicate significant differences among teams in expected manner</td>
<td>N/A (self-report)</td>
<td>High: total scale ( \alpha ) was .94</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenaszchuk et al. (2010)</td>
<td>Adapted version of Nurses’ Opinion Questionnaire (NOQ) of the Ward Organisational Features Scales</td>
<td>Moderate evidence: concurrent validity evidence (e.g., performed pairwise hospital site comparisons of mean scale score for NWI-NPRS among ratings)</td>
<td>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)</td>
<td>N/A (self-report)</td>
<td>High: Cronbach’s ( \alpha ) ranged from .71 to .88 for subscales</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Scale Title</td>
<td>Evidence Type (examples)</td>
<td>Supporting Evidence</td>
<td>Cronbach’s α (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiesewetter &amp; Fischer (2015)</td>
<td>The Teamwork Assessment Scale (TAS)</td>
<td>Moderate evidence: concurrent validity (e.g., correlated with clinical performance) Moderate evidence: EFA conducted and supported theorized structure but could not differentiate between expected two dimensions</td>
<td>Unable to find supporting evidence</td>
<td>Moderate: Cronbach’s α ranged from .67 to .81 for subscales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King et al. (2010)</td>
<td>The Interprofessional Socialization and Valuing Scale (ISVS)</td>
<td>Unable to find supporting evidence</td>
<td>Moderate evidence: PCA conducted and supported theorized structure</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamb et al. (2011)</td>
<td>Multidisciplinary Team Performance Tool</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lazar (1971)</td>
<td>Team Excellence Questionnaire</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lazar (1985)</td>
<td>Factors Influencing Productivity and Excellence of Team Work</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Unavailable supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Scale/Measure</td>
<td>Supporting Evidence</td>
<td>Correlation Evidence</td>
<td>Reliability Evidence</td>
<td>Supporting Evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------------------------------------------------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lichtenstein et al. (1997)</td>
<td>Team Integration Measure</td>
<td>Unable to find</td>
<td>Strong evidence: correlations supported theorized relationships and discriminant validity evidence (e.g., negatively correlated with age)</td>
<td>Some information available: items based on previous scales</td>
<td>N/A (self-report)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyk-Jensen et al. (2014)</td>
<td>Nurse Anesthetists’ Non-Technical Skills (N-ANTS)</td>
<td>Unable to find</td>
<td>Unable to find</td>
<td>Unable to find</td>
<td>Unable to find</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malec et al. (2007)</td>
<td>Mayo High Performance Teamwork Scale (MHPTS)</td>
<td>Unable to find</td>
<td>Strong evidence: significant difference in pre and post training scores following team training and additional evidence via Rasch indicators</td>
<td>Some information available: based on extensive review of the literature</td>
<td>High: Cronbach’s α was .85 for all ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McClane (1992)</td>
<td>Team Assessment Worksheets</td>
<td>Unable to find</td>
<td>Unable to find</td>
<td>N/A (self-report)</td>
<td>Unable to find</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td></td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heinemann &amp; Zeiss (2002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subscale Description</td>
<td>Validity Evidence</td>
<td>Reliability Evidence</td>
<td>Additional Evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The Oxford Non-Technical Skills (NOTECHS)</strong></td>
<td>Strong evidence: significant difference in pre and post training scores following training (multiple studies)</td>
<td>High: $R_{wg}$ was .99</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Group Environment Scale</strong></td>
<td>Strong: concurrent validity evidence (e.g., correlated with technical error)</td>
<td>N/A (self-report)</td>
<td>Moderate: Cronbach’s $\alpha$ ranged from .65 to .87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The Team Climate Assessment Measurement (TCAM)</strong></td>
<td>Unable to find supporting evidence</td>
<td>N/A (self-report)</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interprofessional Attitudes Scale (IPAS)</strong></td>
<td>Strong evidence: CFA conducted and supported theorized structure and additional EFA conducted (responses randomly split)</td>
<td>N/A (self-report)</td>
<td>Moderate: Cronbach’s $\alpha$ ranged from .62 to .92 for subscales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Perception of Interprofessional</strong></td>
<td>Moderate evidence: PCA conducted</td>
<td>Low: Cronbach’s $\alpha$</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Moos (1986) and Heinemann & Zeiss (2002); Littlepage et al. (1989); Moos (1994)
<p>| 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 | High: interrater reliability was .88 | Unable to find supporting evidence | Unable to find supporting evidence |
| Parsell &amp; 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, &amp; Gilchrist (2004) UWE Entry Level Interprofessional Questionnaire, ELIQ | Moderate evidence: concurrent validity evidence (e.g., correlated with IEPS; Leuchter 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) |</p>
<table>
<thead>
<tr>
<th>Authors and Year</th>
<th>Instrument Name and Version</th>
<th>Evidence Quality</th>
<th>Supporting Evidence</th>
<th>Detailed Information Available</th>
<th>Cronbach’s α Ranges</th>
<th>Author(s) and Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roberton et al. (2014)</td>
<td>The Oxford Non-Technical Skills (NOTECHS) II</td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with WHO time-out)</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on expert review, previous measure, and pilot testing</td>
<td>Low: levels of agreement ranged between 45% and 78%</td>
<td>Unable to find supporting evidence</td>
</tr>
<tr>
<td>Rothermich &amp; Saunders (1977)</td>
<td>Team Effectiveness Rating Scale</td>
<td>Unable to find supporting evidence</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on extensive review of the literature, expert review, and previous measures</td>
<td>N/A (self-report)</td>
<td>Unable to find supporting evidence</td>
</tr>
<tr>
<td>Schroder et al. (2011)</td>
<td>Collaborative Practice Assessment Tool (CPAT)</td>
<td>Strong evidence: EFA conducted and CFA conducted and supported theorized structure (multiple studies)</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on extensive review of the literature, expert review, and pilot testing</td>
<td>N/A (self-report)</td>
<td>High: Cronbach’s α ranged from .73 to .84 for subscales</td>
</tr>
<tr>
<td>Shortell et al. (1991)</td>
<td>Intensive Care Unit Nurse/Physician Instrument</td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with communication) and discriminant validity evidence (e.g., negatively correlated with</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on pilot testing</td>
<td>N/A (self-report)</td>
<td>Moderate: Cronbach’s α ranged from .64 to .94 for subscales</td>
</tr>
<tr>
<td>Study</td>
<td>Tool Description</td>
<td>Evidence Details</td>
<td>Reliability Measures</td>
<td>Validity Details</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singleton et al. (1999)</td>
<td>McMaster-Ottawa Team Observed Structured Clinical Encounter (TOSCE)</td>
<td>Unable to find supporting evidence</td>
<td>Moderate</td>
<td>Generalizability study (G-study)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taylor, Atkins et al. (2012)</td>
<td>Multidisciplinary team observational assessment rating scale (MDT-OARS)</td>
<td>Unable to find supporting evidence</td>
<td>Low: ICCs ranged from .32-.92</td>
<td>Cronbach’s α ranged from .52 to .81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taylor, Brown et al. (2012)</td>
<td>Team Evaluation and Assessment Measure (TEAM)</td>
<td>Unable to find supporting evidence</td>
<td>N/A (self-report)</td>
<td>Cronbach’s α ranged from .76 to .89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temkin-Greener et al. (2004)</td>
<td>Interdisciplinary Team Performance Scale (ITPS)</td>
<td>Unable to find supporting evidence</td>
<td>N/A (self-report)</td>
<td>Cronbach’s α was .97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thompson et al. (2009)</td>
<td>Team Performance Scale (TPS)</td>
<td>Unable to find supporting evidence</td>
<td>N/A (self-report)</td>
<td>High: total scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Instrument/Tool</td>
<td>Supporting Evidence</td>
<td>Evidence Quality</td>
<td>N/A (Self-Report)</td>
<td>Unavailable Evidence</td>
<td>Report Authors</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Tsukuda &amp; Stahelski (1990)</td>
<td>Team Skills Questionnaire</td>
<td>Unable to find</td>
<td>Unable to find</td>
<td>N/A (self-report)</td>
<td>Unable to find</td>
<td>Heinemann &amp; Zeiss (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>supporting evidence</td>
<td>supporting evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walker et al. (2011)</td>
<td>Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR)</td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with TEAM measure)</td>
<td>Unable to find supporting evidence</td>
<td>Detailed information available: items based on extensive literature review, expert review, and previous measures</td>
<td>Moderate: ICCs ranged from .61 to .88</td>
<td>Moderate: Cronbach’s α ranged from .74 to .97 for subscales</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wallin et al. (2007)</td>
<td>Emergency medicine crisis resource management (EMCRM)</td>
<td>Unable to find supporting evidence</td>
<td>Moderate evidence: significant</td>
<td>Detailed information available:</td>
<td>Moderate: inter-rater reliability</td>
<td>Unable to find supporting evidence</td>
</tr>
<tr>
<td>Measure</td>
<td>Supporting Evidence</td>
<td>Evidence Type</td>
<td>Cronbach’s $\alpha$ Range</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>---------------------</td>
<td>----------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Weller et al. (2011)</strong></td>
<td>Strong evidence: EFA conducted and significant difference in scores over time</td>
<td>Strong evidence</td>
<td>Range from .60 to .78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modified Version of the Mayo High Performance Teamwork Scale</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wheelan &amp; Hochberger (1993)</strong></td>
<td>Moderate evidence: concurrent validity evidence (e.g., correlated with Group Attitude Scale; Evans &amp; Jarvis, 1986)</td>
<td>Strong evidence</td>
<td>Range from .69 to .82 for subscales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group Development Questionnaire (GDQ)</td>
<td>Strong evidence: scale indicates significant differences among teams in expected manner (multiple studies)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yule et al. (2006)</strong></td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-technical Skills for Surgeons (NOTTs) Rating Scale</td>
<td>Unable to find supporting evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOTE: Some measures excluded because we could not access ANY information about them (Helmreich’s ORMAQ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>