Reliability of Observational Assessment Methods for Outcome-based Assessment of Surgical Skill: Systematic Review and Meta-analyses

Marleen Groenier, PhD,* Leonie Brummer, MSc,* Brendan P. Bunting, PhD,† and Anthony G. Gallagher, PhD, DSc‡

*Technical Medical Centre, University of Twente, Enschede, The Netherlands; †Psychology Research Institute, Ulster University, Coleraine, Northern Ireland; and ‡ASSERT Centre, College of Medicine and Health, University College Cork, Cork, Ireland

BACKGROUND: Reliable performance assessment is a necessary prerequisite for outcome-based assessment of surgical technical skill. Numerous observational instruments for technical skill assessment have been developed in recent years. However, methodological shortcomings of reported studies might negatively impinge on the interpretation of inter-rater reliability.

OBJECTIVE: To synthesize the evidence about the inter-rater reliability of observational instruments for technical skill assessment for high-stakes decisions.

DESIGN: A systematic review and meta-analysis were performed. We searched Scopus (including MEDLINE) and Pubmed, and key publications through December, 2016. This included original studies that evaluated reliability of instruments for the observational assessment of technical skills. Two reviewers independently extracted information on the primary outcome (the reliability statistic), secondary outcomes, and general information. We calculated pooled estimates using multilevel random effects meta-analyses where appropriate.

RESULTS: A total of 247 documents met our inclusion criteria and provided 491 inter-rater reliability estimates. Inappropriate inter-rater reliability indices were reported for 40% of the checklists estimates, 50% of the rating scales estimates and 41% of the other types of assessment instruments estimates. Only 14 documents provided sufficient information to be included in the meta-analyses. The pooled Cohen’s kappa was .78 (95% CI 0.69-0.89, p < 0.001) and pooled proportion agreement was 0.84 (95% CI 0.71-0.96, p < 0.001). A moderator analysis was performed to explore the influence of type of assessment instrument as a possible source of heterogeneity.

CONCLUSIONS AND RELEVANCE: For high-stakes decisions, there was often insufficient information available on which to base conclusions. The use of suboptimal statistical methods and incomplete reporting of reliability estimates does not support the use of observational assessment instruments for technical skill for high-stakes decisions. Interpretations of inter-rater reliability should consider the reliability index and assessment instrument used. Reporting of inter-rater reliability needs to be improved by detailed descriptions of the assessment process. (J Surg Ed 77:189–201. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: outcome-based assessment, surgical skill, inter-rater reliability, reporting guidelines

COMPETENCIES: Patient Care, Medical Knowledge

INTRODUCTION

The “Bristol Case”¹ and the “To Err is Human”² reports revealed a major deficiency in the area of surgical education, training, and assessment. There was no uniform or consistent training in surgical skills, either at a local or national level. Surgical training continued in the traditional mentoring method, where students were exposed to patient care with the guidance of an experienced...
surgeon teacher. The Institute of Medicine in the United States in a report published in July 2014 proposed that Graduate Medical Education must move from a process driven enterprise to one that is “outcome” driven.\textsuperscript{5} Outcome-based assessment means that not only the amount of experience (i.e., time in training, procedures done etc.) should determine progression in training or licensing, but more importantly the demonstration of a predefined level of performance (or milestones). Thus, reliable and valid performance assessment is of increasing importance and moving toward a situation where these assessments involve “high stakes”. Such high-stakes assessments are any evaluations or tests which have important implications for the test taker, for example, a resident or practicing surgeon can progress or may be removed from their training program, or lose his or her practice license. Using measurement instruments in such high-stakes assessments calls for a critical analysis of the validity and reliability of these instruments.\textsuperscript{4}

In the last 2 decades, numerous observational instruments have been developed for technical skill assessment inside and outside the operating room.\textsuperscript{5,8} Reviews\textsuperscript{6,5,10} suggest that these various assessment instruments are reliable and can be used for the evaluation of performance in actual practice. For example, Reznick and MacRae\textsuperscript{11} have suggested that the Objective Structured Assessment of Technical Skill (OSATS) is “acceptable for summative high-stakes evaluation purposes” (p. 2665). However, as Swanson and Van der Vleuten\textsuperscript{12} point out, interpretation of results from these studies may be difficult because of methodological shortcomings which negatively impinge on the interpretation of the results. Validity of an assessment is seriously compromised if an assessment instrument is unreliable. Reliability refers to the consistency of outcomes of an instrument for repeated measurements under several conditions, such as over time or by different observers.\textsuperscript{13} Fundamental to this process is the requisite that observers need to agree on the assessed performance that is scored.

Inter-rater reliability refers to the degree with which 2 or more observers assign the same score to an individual’s performance when using the same assessment instrument.\textsuperscript{14,15} It is crucial that measures used to evaluate inter-rater reliability should take into account the extent to which observers assign the same scores to a trainee’s performance. Acceptable measures for determining inter-rater reliability are therefore those based on agreement, such as Cohen’s kappa.\textsuperscript{16,17} Statistical measures such as Cronbach’s alpha or the correlation coefficient are inappropriate for evaluating inter-rater agreement because they are measures of association and not agreement.\textsuperscript{16–18} Cronbach’s alpha relies on the correlations between scores on individual items of the test and is therefore a measure of association, not agreement. The limitation of inter-rater reliability measures based on association is that the association between the scores of 2 different observers can be perfect, even though they disagree on every item they scored.\textsuperscript{19} Therefore, one needs to take into account the type of inter-rater reliability index that was used when making a statement about the reliability of an assessment instrument as the interpretation will depend on the underlying assumptions of each approach.

According to international standards,\textsuperscript{20} it is contended that an assessment instrument should meet 2 requirements of inter-rater reliability to be used in high-stakes assessments: (1) inter-rater reliability should be at least 0.90\textsuperscript{21} and (2) this reliability should be based on the amount of agreement between the observers.\textsuperscript{22} The purpose of this review was to critically appraise and compare the evidence on the inter-rater reliability of various observational assessment instruments for the evaluation of technical surgical skill. To this end, a qualitative systematic review was performed and complemented with meta-analyses to synthesize research outcomes and examine factors influencing inter-rater reliability. Based on these analyses, an evaluation is made of assessment instruments which could meet the requirements for high-stakes decisions.

**METHOD**

**Search**

We searched Scopus, including MEDLINE, and PUBMED until December 2016 for relevant peer reviewed manuscripts published in English about technical surgical skill assessment. The first (MG) and last (AG) author determined the search strategy, the first author (MG) performed the search. Duplicates were identified by the Endnote reference manager program as well as manually by MG. There is no registered protocol for the systematic review, but Supplementary Material 1 (SM1) contains the full search strategy used. To identify published studies further, we cross checked the reference lists from the recent systematic reviews for the objective assessment of technical skill by Van Hove et al.\textsuperscript{6} and Ahmed et al.\textsuperscript{10} with the documents retrieved in the initial search.

**Study Selection**

The results from the literature search were screened by the first (MG) and last (AG) author independently by reading the title and/or abstract. To gain as many relevant studies as possible we determined broad inclusion criteria:

1. Original research studies using a measure of inter-rater reliability to evaluate technical skill assessment by means of either direct or video observation;
2. Participants with various experience levels (from medical student to expert);
3. Assessors with various experience levels (from medical student to expert);
4. Studies reporting on any type of surgical skill or procedure, including both open and image-guided procedures, from any specialty;
5. Studies reporting on assessments made in simulated environments or in the operating theatre.

Only documents that reported overall reliability estimates were included. Reliability estimates at the level of specific items of the assessment instrument or for different stations in an examination (i.e., different tasks/procedures are assessed) were not considered overall estimates and therefore excluded. Multiple overall reliability estimates could be reported in the same document. An overall estimate was defined as an estimate for:

1. A specific type of assessment instrument, for example, a reliability estimate was reported for both the checklist and the global rating scale of an OSATS;
2. A specific group of participants, for example, separate reliability estimates were calculated for medical students and residents;
3. A subgroup of participants used to calculate an overall score, for example, separate reliability estimates for both the complete sample as well as for a particular subset of participants;
4. A subgroup of assessors and/or different numbers of assessors, for example, separate reliability estimates for both experienced and inexperienced assessors.

Exclusion criteria were:

1. Studies on team assessment or training, communication, patient management, physical examination, and/or nontechnical skills;
2. Studies assessing technical skills of dentists, veterinarians, and/or nurses;
3. Retrospective study designs, reviews, editorials, letters, and notes;
4. Studies using data from records (e.g., ward evaluations at the end of an internship).

**Data Extraction**

Data from included documents were extracted using a data extraction sheet with variables about general information, primary outcomes, and secondary outcomes, see SM2 for an overview of all variables. To assess risk of bias and methodological quality we extracted data regarding the training and blinding of assessors, participant and assessor demographics, and the assessment situation, see SM2. Intercoder agreement was determined in 2 stages.

First, the titles and abstracts were divided into groups of 50 and randomly allocated to the first (MG) or last (AG) author to review. From each of these groups, 5 titles and abstracts were randomly selected and independently checked by the other author to calculate intercoder agreement. This resulted in a sample of 84 randomly selected titles and abstracts reviewed for inclusion by the first (MG) and last (AG) author independently to establish intercoder agreement. Proportion agreement ($p_a = \frac{\text{number of agreements}}{\text{total number of documents selected}}$) for including a document was 1.0.

Second, data from the included documents were extracted by the first (MG) and second (LB) author independently. Three to seven rounds of data extraction and discussion about the differences in coding were necessary to achieve acceptable intercoder agreement. A total of 82 additional documents were randomly selected in the 7 rounds to evaluate intercoder agreement. Cohen’s kappa’s (SE) were calculated for categorical variables, and 2-way mixed effects single measures absolute agreement IntraClass Correlation (ICC) coefficients (95% CI) were calculated for ordinal or continuous variables, see SM2.

**Methodological Quality Assessment**

Several aspects of an assessment situation influence reliability. Participant and assessor characteristics, such as the number of participants, assessor training and experience level influence the magnitude of the interrater reliability estimate. In addition, information about statistical uncertainty, such as confidence intervals or standard errors, is crucial to interpretation of the precision of measurement. A qualitative analysis of study quality was therefore performed by examining characteristics of participants and assessors, description of the assessment process, and reporting of statistical uncertainty measures.

**Synthesis and Statistical Analysis**

Overall inter-rater reliability of surgical skill assessment was analyzed qualitatively and quantitatively based on the type of (1) assessment instrument that was used and (2) reliability index reported. To facilitate analysis and interpretation of the results the assessment instruments were grouped into 3 categories: (1) procedure-specific checklists, (2) rating scales, and (3) other assessment instruments, for example, pass/fail decisions, final result assessments, and visual-analog scales. The main difference between procedure-specific checklists and rating scales is the response format. Whereas the response format of a procedure-specific checklist is dichotomous
(yes/no), the response format of both a procedure-specific and a global rating scale is more elaborate, such as a 5- or 10-point scale, often ranging from “unsatisfactory” to “excellent”. We combined procedure-specific and global rating scales in the analysis because they share a common response format.

Furthermore, the inter-rater reliability indices were grouped into 3 categories: (1) association-based indices (e.g., correlation coefficient, Cronbach’s alpha coefficient), (2) agreement-based indices (e.g., Cohen’s kappa, proportion agreement), and (3) other indices (e.g., Kendall’s tau, British Standard Institution Reproducibility Coefficient, generalizability theory). Reliability estimates with missing information about the type of reliability index or assessment instrument used were excluded.

Meta-Analysis
Quantitative analysis consisted of meta-analysis to pool inter-rater reliability coefficients and apply meta-analytic techniques to synthesize research outcomes and explore sources of heterogeneity. Separate meta-analyses were performed for each type of inter-rater reliability index. In the current analysis, multilevel random effects models were used because both within- and between-study variability can then be taken into account. Residual heterogeneity was assessed by examining the tests for residual heterogeneity.

For the meta-analyses of Cohen’s kappa and proportion agreement the estimates and standard errors were extracted or calculated based on the available information in the documents. Cohen’s kappa estimates were pooled using the procedure described by Sun. There are several types of ICC, see Shrout and Fleiss and McGraw and Wong. For the current analysis the ICC (A,1) would be suitable because this type of ICC provides information about a single rater and takes systematic differences between raters into account. Other types of the ICC provide information about averages of multiple raters or are based on correlations between scores (they are association-based) and are therefore not appropriate to determine inter-rater reliability. The ICC (A,1) is also often described as a 2-way mixed effects single measures absolute agreement ICC. However, to our knowledge there is currently no statistical technique available to calculate the standard error or variance for this type of ICC, and for this reason a meta-analysis has not also been conducted.

Some documents reported more than 1 overall inter-rater reliability estimate, for example, for both a checklist and a rating scale, which resulted in dependent estimates. Dependent observations cause bias in the estimation of the pooled reliability estimates; therefore, we applied multilevel random effects meta-analytic techniques. Moderator analyses were performed for procedure-specific checklists, rating scales, and other types of instruments. The multilevel random effects meta-analyses were fitted using R package metafor (https://www.r-project.org/). Descriptive statistical analyses were performed with SPSS (version 22.0).

RESULTS
Search and Selection of Studies
The PRISMA guidelines were followed during the search and selection of documents, see SM3. The search identified 3307 unique documents, which were assessed for relevance. A total of 718 full text documents were reviewed and 229 documents were excluded. Of the remaining 489 documents, 247 documents met the inclusion criteria, see Figure 1.

Characteristics of the Included Studies
Most documents (n = 118; 48%) reported enrolling participants with varying levels of experience (e.g., a sample consisting of medical students and residents). In 15 documents the number of participants enrolled could not be determined. In 152 documents (62%) participants’ surgical skill performance was assessed in a simulated environment with 89 documents reporting assessment of an image-guided skill in a simulated environment. In 2 documents the type of assessment situation could not be determined. Participants performed various surgical tasks, such as laparoscopic suturing, dissection, and salpingectomy. Consultants (e.g., staff, faculty, fellows) were most often reported as assessors (n = 76; 31%).

Analysis of Methodological and Reporting Quality
Of the 247 documents, 15 (6%) failed adequately to report the number of participants providing data. Whether assessors were trained prior to the actual assessment could not be determined in almost 2 thirds of the documents (64%) and in 62 documents (25%) the use of trained assessors was reported. In addition, 16 documents (6%) failed to report the number of assessors adequately. In about one quarter of the documents (n = 64) the assessor’s experience could not be determined clearly. Furthermore, blinding of assessors to participants’ identities and training levels is important to reduce biased assessments. In 152 documents (62%) blinded assessors were used. In 74 documents (30%) it was unclear whether assessors were blinded or not. In 78% of the documents, information regarding statistical uncertainty was not reported or could not be determined clearly.
FIGURE 1. PRISMA flow diagram for the selection of documents.
Qualitative Analysis of Inter-Rater Reliability

Assessment Instruments
A total of 491 inter-rater reliability estimates were reported in the 247 documents (mean = 2.0; mode = 1; range = 1-18). The majority of documents reported 1 or 2 overall estimates (79%). The Table in SM4 summarizes the number of documents reporting overall reliability estimates for each assessment instrument and reliability index category. In most documents (n = 155; 63%) reliability estimates for 1 assessment instrument category were reported, most often for rating scales (n = 155; 61%). Association-based inter-rater reliability estimates were most often reported for all 3 assessment instrument categories. It should be noted that 6 documents (3%) reported both association- and agreement-based estimates.

Association- Versus Agreement-Based Reliability
A total of 420 association- and agreement-based reliability estimates reported in 220 documents were examined further. Estimates from the category “other types of reliability indices” were excluded because some of these estimates exceeded the range of 0-1 (n = 71). About half of the remaining 420 estimates were based on association-based reliability indices which are inappropriate to determine inter-rater reliability. The association-based indices correlation and Cronbach’s alpha were used to determine inter-rater reliability for 40%, 50%, and 41% of the checklists, rating scales, and other instruments respectively. In Figure 2 the distribution of only the agreement-based estimates (n = 255; 53%), including the ICC, is presented.

It shows that the ICC, irrespective of the type of ICC, is used most often to determine inter-rater reliability for rating scales. Also, more estimates are 0.90 or higher, the criterion for the reliability of high-stakes assessments, for checklists compared to rating scales. None of the Cohen’s kappa and proportion agreement estimates reached 0.90 for the rating scales. The number of reported estimates based on an inappropriate measure (i.e., association) is even higher if the ICC is considered an association based index: 77%, 92%, and 79% for checklists, rating scales, and other instruments respectively.

Meta-Analysis of Inter-Rater Reliability
For the quantitative analysis, we included those agreement-based estimates for which the necessary information to perform the meta-analysis could be retrieved or calculated from the documents (N = 21), see Figure 3. The study characteristics are given in Table 1.

As can be seen in Table 1, the studies differed in a number of ways. In 10 documents the use of a procedure-specific checklist was used, in 5 documents a rating scale and in 4 documents a pass/fail decision was used. The included studies not only differed in the method of assessment but also in the reliability index used. Furthermore, the studies differed in the type of participants and
raters used. Residents were most often assessed \( n = 6 \) while consultants were most often raters \( n = 7 \).

To take this within- and between study variability into account, we used a multilevel random effects meta-analysis model and explored heterogeneity. We expected that the type of assessment instrument used would most likely influence the magnitude of the reliability estimate. Therefore, we also fitted random effects
<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Assessment Instrument</th>
<th>Reliability Index</th>
<th>Assessment Situation</th>
<th>Participants</th>
<th>Sample Size</th>
<th>Assessors</th>
<th>Training</th>
<th>Blinding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure-specific checklists</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seymour NE36</td>
<td>2002</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>In vivo/image-guided</td>
<td>Residents</td>
<td>16</td>
<td>Consultants</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sarker SK37</td>
<td>2005</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>In vivo/image-guided</td>
<td>Consultants</td>
<td>8</td>
<td>Consultants</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Ahlberg G38</td>
<td>2007</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>Simulated/image-guided</td>
<td>Residents</td>
<td>13</td>
<td>Experts</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Laeeq K39</td>
<td>2009</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>23</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Gallagher AG19</td>
<td>2014</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>19</td>
<td>Consultant</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Andersen SA40</td>
<td>2015</td>
<td>Task-specific checklist</td>
<td>Cohen’s kappa</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>34</td>
<td>Experts</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Wong IH41</td>
<td>2014</td>
<td>Task-specific checklist</td>
<td>Cohen’s kappa</td>
<td>Simulated/image-guided</td>
<td>Medical students</td>
<td>35</td>
<td>Consultant</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Angelo RL42</td>
<td>2015</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>Simulated/image-guided</td>
<td>Mixed</td>
<td>19</td>
<td>Consultant</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Angelo RL43</td>
<td>2015</td>
<td>Task-specific checklist</td>
<td>Proportion agreement</td>
<td>Simulated/image-guided</td>
<td>Mixed</td>
<td>22</td>
<td>Consultant</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Day RW44</td>
<td>2016</td>
<td>Task-specific checklist</td>
<td>Cohen’s kappa</td>
<td>Simulated/open</td>
<td>Mixed</td>
<td>41</td>
<td>Mixed</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Rating scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laeeq K39</td>
<td>2009</td>
<td>Global rating scale</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>23</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Fried MP45</td>
<td>2010</td>
<td>Global rating scale</td>
<td>Cohen’s kappa</td>
<td>Combination</td>
<td>Residents</td>
<td>25</td>
<td>Experts</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Gallagher AG19</td>
<td>2014</td>
<td>OSATS global rating scale</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Unknown</td>
<td>19</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Wong IH41</td>
<td>2015</td>
<td>Global rating scale</td>
<td>Cohen’s kappa</td>
<td>Simulated/image-guided</td>
<td>Medical students</td>
<td>35</td>
<td>Consultant</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Iordache F46</td>
<td>2015</td>
<td>Task-specific rating scale</td>
<td>Cohen’s kappa</td>
<td>Simulated/image-guided</td>
<td>Mixed</td>
<td>20</td>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>Other instruments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laeeq K39</td>
<td>2009</td>
<td>Pass/fail decision</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>23</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Ma IW47</td>
<td>2012</td>
<td>Pass/fail decision</td>
<td>Proportion agreement</td>
<td>Simulated/open</td>
<td>Residents</td>
<td>23</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Koehler RJ48</td>
<td>2013</td>
<td>Pass/fail decision</td>
<td>Proportion agreement</td>
<td>Simulated/image-guided</td>
<td>Mixed</td>
<td>30</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Wong IH41</td>
<td>2015</td>
<td>Pass/fail decision</td>
<td>Cohen’s kappa</td>
<td>Simulated/image-guided</td>
<td>Medical students</td>
<td>35</td>
<td>Consultant</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
</tbody>
</table>
models for Cohen’s kappa and proportion agreement with the assessment instrument category as a moderator. Results from the meta-analyses are reported in Table 2.

The pooled Cohen’s kappa and proportion agreement for the models without the assessment instruments as moderators were 0.78 and 0.84 respectively, indicating substantial agreement between assessors. Random effects models were also fitted with the assessment instrument category included as a moderator. The pooled Cohen’s kappa was lowest for the pass/fail decisions and comparable for the procedure-specific checklists and the rating scales. The pooled proportion agreement was highest for pass/fail decisions and lowest for rating scales.

The tests for heterogeneity were significant for both meta-analyses, taking the effect of the different assessment instrument categories into account. $Q_H$ was 75.53 (df = 7, p < 0.0001; $I^2 = 91\%$) for the analysis of Cohen’s kappa and 2870.94 (df = 8, p < 0.0001; $I^2 = 99\%$) for the analysis of proportion agreement. This indicates that other moderators not considered in the models were influencing inter-rater reliability.

**DISCUSSION**

Graduate medical education is moving toward an “outcome” driven approach where trainees are required to demonstrate a predefined level of technical skill performance before progressing in training. Evaluation of performance is crucial to provide feedback to the trainee, as well as ensuring that a trainee sufficiently masters a skill for independent practice. What constitutes a valid and reliable assessment instrument is a well-established discussion in the behavioral sciences and has resulted in international standards for testing. Application of these standards in medical education research and practice has not been consistent.

As stated above, an assessment instrument should meet 2 requirements of inter-rater reliability to be used in high-stakes assessments: (1) inter-rater reliability should be at least 0.90 and (2) this reliability should be based on the amount of agreement between the observers rather than the amount of association between the scores. Only 14% of the reported inter-rater reliability estimates in our review were above 0.90 and based on agreement (including the ICC). Also, a substantial amount of the documents lacked information necessary to summarize the information in a meta-analysis statistically. This resulted in a marked reduction of the number of documents that could be included in our meta-analysis: only 14 out of 247 documents.

Based on this analysis, considerable caution is required before the use of many of these assessment instruments, at least where high-stake decision making is required. Sub-optimal methods to determine inter-rater reliability in combination with incomplete reporting of inter-rater reliability evaluations prohibiting valid judgement about the reliability of observational assessment instruments for technical skill were often evident. However, there is abundant reliability evidence supporting the use of these instruments in formative assessment aimed at providing feedback to learners, see, for example, the reviews by Van Hove et al.6 and Ahmed et al.10 and the meta-analysis of OSATS by Hatala et al.23 The current study adds to these previous reviews by identifying problems in the published literature with the design and reporting of reliability studies.

**Limitations of Evidence**

Both the qualitative and quantitative evaluation of inter-rater reliability showed that reliability for rating scales was generally lower than for checklists or other types of instruments. However, these results should be interpreted with caution. Given the nature of the data, the analysis of model heterogeneity was problematic. A number of factors made it difficult to evaluate statistically the inter-rater reliability of observational assessment instruments. Information about sample selection, study design, statistical analysis, and information relating to the reliability estimates statistically was often incomplete or ambiguous. Comparison across diverse methods of assessment is likely to contain substantial method effects, and in the current study these differential effects are illustrated. We therefore cannot conclude that, for example, the use of checklists results in higher inter-rater reliability than rating scales, because this depends

---

**TABLE 2.** Pooled Inter-Rater Reliability Estimates and Confidence Intervals (CI) for Multilevel Random Effects Regression Models for Cohen’s Kappa and Proportion Agreement

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>Pooled Estimate</th>
<th>CI</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No moderators</td>
<td>7</td>
<td>0.78</td>
<td>0.69-0.89</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: checklists</td>
<td>4</td>
<td>0.82</td>
<td>0.69-0.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: rating scales</td>
<td>3</td>
<td>0.79</td>
<td>0.63-0.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: other instruments</td>
<td>2</td>
<td>0.61</td>
<td>0.37-0.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No moderators</td>
<td>6</td>
<td>0.84</td>
<td>0.71-0.96</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: checklists</td>
<td>5</td>
<td>0.84</td>
<td>0.72-0.97</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: rating scales</td>
<td>2</td>
<td>0.69</td>
<td>0.52-0.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moderator: other instruments</td>
<td>2</td>
<td>1.0</td>
<td>0.84-1.2</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
on many other factors, such as the reliability index used, the assessment situation (e.g., in vivo or simulation), the procedure that is performed, and the experience level of participants and raters.

We found that association- and agreement-based reliability indices are reported equally often, and we also noted similar interpretations of inter-rater reliability estimates irrespective of the type of reliability index used. Association-based reliability indices, such as the correlation and Cronbach’s alpha coefficient, have the disadvantage that they imply that a relationship between scores exists, merely assessing the extent to which scores go together. The best approach to evaluate inter-rater reliability is to analyze systematic differences and chance agreement between assessors which necessitates the use of agreement-based indices, such as Cohen’s kappa.22

**Guidelines for the Reporting of Inter-Rater Reliability**

We describe guidelines for reporting statistical information of inter-rater reliability evaluation studies. These guidelines are aimed at improving reporting practices so that research results from inter-rater reliability studies can be aggregated and analyzed. For general reporting guidelines of inter-rater reliability studies we refer to Kottner et al.24

1. Specify the subject population of interest: number of participants used for inter-rater reliability evaluation, participants’ level of experience, and demographics.
2. Specify the assessor population of interest: number of assessors, assessors’ level of experience, and demographics.
3. Describe the assessment process: blinding and training of assessors, how assessors were assigned to participants (was the design fully crossed? See Hallgren).15
4. State the number of replicate observations.
5. State which reliability index was used to evaluate inter-rater reliability. Report inter-rater agreement rather than inter-rater consistency or association.
   a. Percentage or proportion agreement: report (i) the estimate, (ii) the sample size, and (iii) the number of observations per participant.
   b. Cohen’s kappa: report (i) the estimate, (ii) the percentage or proportion agreement, (iii) the sample size, and (iv) the number of observations per participant.
   c. ICC: report (i) the type of ICC according to the classification by McGraw and Wong,34 (ii) the estimate, (iii) the sample size, and (iv) the number of observations per participant.
6. Provide information about the statistical precision of measurement. Report either a standard error or a confidence interval.

**Strengths and Limitations**

The strengths of the current study are that we included a broad range of studies reporting about various surgical specialties and assessment situations; while (1) critically analyzing the methods used to evaluate inter-rater reliability, (2) distinguishing between different types of inter-rater reliability indices and (3) evaluating their appropriateness for the intended purpose. We provide specific examples of meta-analytic techniques applied to reliability studies. Furthermore, we present guidelines for reporting inter-rater reliability studies to improve reporting practice, thereby enabling future work on aggregating reliability evidence for observational assessment of technical skill.

A limitation is that only overall estimates were included. Documents that reported separate estimates for performance assessment in different situations (e.g., operating room vs. bench model), for different procedures, or for each item of an instrument were excluded. Also, our analysis was focused on inter-rater reliability, and in follow-up studies we will examine other types of reliability. Finally, every attempt was made to minimize selection bias. However, there is a possibility that some published studies may not have come to light despite an extensive search of the relevant literature.

**CONCLUSION**

In summary, the evidence for the inter-rater reliability of observational technical skill assessment instruments for high-stakes decisions is inconclusive. Although many studies report substantial to high inter-rater reliability for a variety of instruments, these studies should be interpreted with caution because of the use of suboptimal methods to evaluate inter-rater reliability. Furthermore, we identified several problems with the reporting of statistical information in the majority of published studies on inter-rater reliability. We present guidelines for the reporting of inter-rater reliability studies to encourage accurate reporting of statistical information thereby enabling the statistical aggregation of evidence in the future.

**ACKNOWLEDGMENTS**

No preregistration exists for the study reported in this article. We would like to thank Hanneke Becht and Marjolein Drent from the University of Twente for their help with establishing the search strategy.
REFERENCES


3. Institute of Medicine. Committee on the Governance and Financing of Graduate Medical Education. Graduate Medical Education that meets the Nation's Health Needs. Institute of Medicine: Washington, DC; 2014.


**SUPPLEMENTARY INFORMATION**

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2019.07.007.