



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

Marlene, G., Brummer, L., Bunting, B., & Gallagher, A. G. (2019). Reliability of Observational Assessment Methods for Outcome-based Assessment of Surgical Skill: Systematic Review and Meta-analyses. *Journal of Surgical Education*, Article 2031. Advance online publication. <https://doi.org/10.1016/j.jsurg.2019.07.007>

[Link to publication record in Ulster University Research Portal](#)

Published in:
Journal of Surgical Education

Publication Status:
Published online: 20/08/2019

DOI:
[10.1016/j.jsurg.2019.07.007](https://doi.org/10.1016/j.jsurg.2019.07.007)

Document Version
Author Accepted version

General rights

The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at <https://creativecommons.org/share-your-work/licenses/>.

Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

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

Marleen Groenier PhD ^a, Leonie Brummer MSc ^a, Brendan P. Bunting PhD ^b,
Anthony G. Gallagher PhD, DSc ^c

^a Department of Technical Medicine, University of Twente, Enschede, The Netherlands

^b Psychology Research Institute, Ulster University, Coleraine, Northern Ireland

^c ASSERT Centre, College of Medicine and Health, University College Cork, Cork, Ireland

Word count: approx. 4000

Declarations of interest: none

Funding: This research did not receive any specific grant from funding agencies in the
public, commercial, or not-for-profit sectors.

Brief title: Reliability of outcome-based assessment

Address correspondence or requests for reprints to Marleen Groenier, Faculty of Science and
Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; E-mail:
m.groenier@utwente.nl; Phone: +31 53 4895569; Fax: +31 53 489 3288

27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

Abstract

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 .69-.89, $p < .001$) and pooled proportion agreement was .84 (95% CI .71-.96, $p < .001$). 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.

Keywords: outcome-based assessment; surgical skill; inter-rater reliability; reporting guidelines

ACGME competences: patient care; medical knowledge

55

Introduction

56 The 'Bristol Case' ¹ and the 'To Err is Human' ² reports revealed a major deficiency in the
57 area of surgical education, training, and assessment. There was no uniform or consistent training in
58 surgical skills, either at a local or national level. Surgical training continued in the traditional
59 mentoring method, where students were exposed to patient care with the guidance of an experienced
60 surgeon teacher. The Institute of Medicine in the USA in a report published in July 2014 proposed that
61 Graduate Medical Education must move from a process driven enterprise to one that is 'outcome'
62 driven ³. Outcome-based assessment means that not only the amount of experience (i.e., time in
63 training, procedures done etc.) should determine progression in training or licensing, but more
64 importantly the demonstration of a predefined level of performance (or milestones). Thus, reliable and
65 valid performance assessment is of increasing importance and moving towards a situation where these
66 assessments involve 'high-stakes'. Such high-stakes assessments are any evaluations or tests which
67 have important implications for the test taker, e.g., a resident or practicing surgeon can progress or
68 may be removed from their training program, or lose his or her practice license. Using measurement
69 instruments in such high-stakes assessments calls for a critical analysis of the validity and reliability of
70 these instruments ⁴.

71 In the last two decades, numerous observational instruments have been developed for
72 technical skill assessment inside and outside the operating room (OR) ⁵⁻⁸. Reviews ^{6,9,10} suggest that
73 these various assessment instruments are reliable and can be used for the evaluation of performance in
74 actual practice. For example, Reznick and MacRae ¹¹ have suggested that the Objective Structured
75 Assessment of Technical Skill is 'acceptable for summative high-stakes evaluation purposes' (p.
76 2665). However, as Swanson and Van der Vleuten ¹² point out, interpretation of results from these
77 studies may be difficult because of methodological shortcomings which negatively impinge on the
78 interpretation of the results. Validity of an assessment is seriously compromised if an assessment
79 instrument is unreliable. Reliability refers to the consistency of outcomes of an instrument for repeated
80 measurements under several conditions, such as over time or by different observers ¹³. Fundamental to
81 this process is the requisite that observers need to agree on the assessed performance that is scored.

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

95 According to international standards ²⁰, it is contended that an assessment instrument should
96 meet two requirements of inter-rater reliability to be used in high-stakes assessments: 1) inter-rater
97 reliability should be at least .90 ²¹ and 2) this reliability should be based on the amount of agreement
98 between the observers ²². The purpose of this review was to critically appraise and compare the
99 evidence on the inter-rater reliability of various observational assessment instruments for the
100 evaluation of technical surgical skill. To this end, a qualitative systematic review was performed and
101 complemented with meta-analyses to synthesize research outcomes and examine factors influencing
102 inter-rater reliability. Based on these analyses, an evaluation is made of assessment instruments which
103 could meet the requirements for high-stakes decisions.

104 Method

105 Search

106 We searched Scopus, including MEDLINE, and PUBMED until December 2016 for relevant
107 peer reviewed manuscripts published in English about technical surgical skill assessment. The first
108 (MG) and last (AG) author determined the search strategy, the first author (MG) performed the search.
109 Duplicates were identified by the Endnote reference manager program as well as manually by MG.

110 There is no registered protocol for the systematic review, but Supplementary Material 1 (SM1)
111 contains the full search strategy used. To identify published studies further, we cross-checked the
112 reference lists from the recent systematic reviews for the objective assessment of technical skill by
113 Van Hove et al. ⁶ and Ahmed et al. ¹⁰ with the documents retrieved in the initial search.

114 **Study selection**

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

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

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

- 130 1. A specific type of assessment instruments, e.g., a reliability estimate was reported for both the
131 checklist and the global rating scale of an Objective Structured Assessment of Technical Skill
132 (OSATS);
- 133 2. A specific group of participants, e.g., separate reliability estimates were calculated for medical
134 students and residents;
- 135 3. A subgroup of participants used to calculate an overall score, e.g., separate reliability estimates for
136 both the complete sample as well as for a particular subset of participants;

137 4. A subgroup of assessors and/or different numbers of assessors, e.g., separate reliability estimates
138 for both experienced and inexperienced assessors.

139 Exclusion criteria were:

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

145 **Data extraction**

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

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

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

165 **Methodological quality assessment**

166 Several aspects of an assessment situation influence reliability²³. Participant and assessor
167 characteristics, such as the number of participants²⁴, assessor training²⁵⁻²⁸ and experience level²⁹
168 influence the magnitude of the inter-rater reliability estimate. In addition, information about statistical
169 uncertainty, such as confidence intervals or standard errors, is crucial to interpretation of the precision
170 of measurement³⁰. A qualitative analysis of study quality was therefore performed by examining
171 characteristics of participants and assessors, description of the assessment process, and reporting of
172 statistical uncertainty measures.

173 **Synthesis and statistical analysis**

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

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

190 **Meta-analysis**

191 Quantitative analysis consisted of meta-analysis to pool inter-rater reliability coefficients and
192 apply meta-analytic techniques to synthesize research outcomes and explore sources of heterogeneity

193 ³¹. Separate meta-analyses were performed for each type of inter-rater reliability index. In the current
194 analysis, multilevel random effects models were used because both within- and between-study
195 variability can then be taken into account. Residual heterogeneity was assessed by examining the tests
196 for residual heterogeneity.

197 For the meta-analyses of Cohen's kappa and proportion agreement the estimates and standard
198 errors were extracted or calculated based on the available information in the documents. Cohen's
199 kappa estimates were pooled using the procedure described by Sun ³². There are several types of ICC,
200 see Shrout and Fleiss ³³ and McGraw and Wong ³⁴. For the current analysis the ICC(A,1) would be
201 suitable because this type of ICC provides information about a single rater and takes systematic
202 differences between raters into account. Other types of the ICC provide information about averages of
203 multiple raters or are based on correlations between scores (they are association-based) and are
204 therefore not appropriate to determine inter-rater reliability. The ICC(A,1) is also often described as a
205 two-way mixed effects single measures absolute agreement ICC. However, to our knowledge there is
206 currently no statistical technique available to calculate the standard error or variance for this type of
207 ICC, and for this reason a meta-analysis has not also been conducted.

208 Some documents reported more than one overall inter-rater reliability estimate, e.g., for both a
209 checklist and a rating scale, which resulted in dependent estimates. Dependent observations cause bias
210 in the estimation of the pooled reliability estimates; therefore, we applied multilevel random effects
211 meta-analytic techniques. Moderator analyses were performed for procedure-specific checklists, rating
212 scales, and other types of instruments. The multilevel random-effects meta-analyses were fitted using
213 *R* package *metafor* ³⁵ (<https://www.r-project.org/>). Descriptive statistical analyses were performed with
214 SPSS (version 22.0).

215 Results

216 Search and selection of studies

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

221 <Insert Figure 1 about here>

222 **Characteristics of the included studies**

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

231 **Analysis of methodological and reporting quality**

232 Of the 247 documents, 15 (6%) failed adequately to report the number of participants
233 providing data. Whether assessors were trained prior to the actual assessment could not be determined
234 in almost two thirds of the documents (64%) and in 62 documents (25%) the use of trained assessors
235 was reported. In addition, 16 documents (6%) failed to report the number of assessors adequately. In
236 about one quarter of the documents (n = 64) the assessor's experience could not be determined clearly.
237 Furthermore, blinding of assessors to participants' identities and training levels is important to reduce
238 biased assessments. In 152 documents (62%) blinded assessors were used. In 74 documents (30%) it
239 was unclear whether assessors were blinded or not. In 78% of the documents, information regarding
240 statistical uncertainty was not reported or could not be determined clearly.

241 **Qualitative analysis of inter-rater reliability**

242 **Assessment instruments**

243 A total of 491 inter-rater reliability estimates were reported in the 247 documents (mean =
244 2.0; mode = 1; range = 1-18). The majority of documents reported one or two overall estimates (79%).
245 The Table in SM4 summarizes the number of documents reporting overall reliability estimates for
246 each assessment instrument and reliability index category. In most documents (n = 155; 63%)
247 reliability estimates for one assessment instrument category were reported, most often for rating scales
248 (n = 155; 61%). Association-based inter-rater reliability estimates were most often reported for all

249 three assessment instrument categories. It should be noted that six documents (3%) reported both
250 association- and agreement-based estimates.

251 **Association- versus agreement-based reliability**

252 A total of 420 association- and agreement-based reliability estimates reported in 220
253 documents were examined further. Estimates from the category ‘other types of reliability indices’ were
254 excluded because some of these estimates exceeded the range of 0 – 1 (n = 71). About half of the
255 remaining 420 estimates were based on association-based reliability indices which are inappropriate to
256 determine inter-rater reliability²². The association-based indices correlation and Cronbach’s alpha
257 were used to determine inter-rater reliability for 40%, 50%, and 41% of the checklists, rating scales,
258 and other instruments respectively. In Figure 2 the distribution of only the agreement-based estimates
259 (n = 255; 53%), including the ICC, is presented.

260 <Insert Figure 2 about here>

261 It shows that the ICC, irrespective of the type of ICC, is used most often to determine inter-
262 rater reliability for rating scales. Also, more estimates are .90 or higher, the criterion for the reliability
263 of high stakes assessments²¹, for checklists compared to rating scales. None of the Cohen’s kappa and
264 proportion agreement estimates reached .90 for the rating scales. The number of reported estimates
265 based on an inappropriate measure (i.e., association) is even higher if the ICC is considered an
266 association based index: 77%, 92%, and 79% for checklists, rating scales, and other instruments
267 respectively.

268 **Meta-analysis of inter-rater reliability**

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

272 <Insert Figure 3 about here>

273 <Insert Table 1 about here>

274 As can be seen in Table 1, the studies differed in a number of ways. In 10 documents the use
275 of a procedure-specific checklist was used, in 5 documents a rating scale and in 4 documents a
276 pass/fail decision was used. The included studies not only differed in the method of assessment but

277 also in the reliability index used. Furthermore, the studies differed in the type of participants and raters
278 used. Residents were most often assessed ($n = 6$) while consultants were most often raters ($n = 7$).

279 To take this within- and between study variability into account, we used a multilevel random
280 effects meta-analysis model and explored heterogeneity. We expected that the type of assessment
281 instrument used would most likely influence the magnitude of the reliability estimate. Therefore, we
282 also fitted random effects models for Cohen's kappa and proportion agreement with the assessment
283 instrument category as a moderator. Results from the meta-analyses are reported in Table 2.

284 <Insert Table 2 about here>

285 The pooled Cohen's kappa and proportion agreement for the models without the assessment
286 instruments as moderators were .78 and .84 respectively, indicating substantial agreement between
287 assessors. Random effects models were also fitted with the assessment instrument category included as
288 a moderator. The pooled Cohen's kappa was lowest for the pass/fail decisions and comparable for the
289 procedure-specific checklists and the rating scales. The pooled proportion agreement was highest for
290 pass/fail decisions and lowest for rating scales.

291 The tests for heterogeneity were significant for both meta-analyses, taking the effect of the
292 different assessment instrument categories into account. QE was 75.53 ($df = 7, p < .0001$) for the
293 analysis of Cohen's kappa and 2870.94 ($df = 8, p < .0001$) for the analysis of proportion agreement.
294 This indicates that other moderators not considered in the models were influencing inter-rater
295 reliability.

296

297

Discussion

298 Graduate medical education is moving towards an 'outcome' driven approach where trainees
299 are required to demonstrate a predefined level of technical skill performance before progressing in
300 training. Evaluation of performance is crucial to provide feedback to the trainee, as well as ensuring
301 that a trainee sufficiently masters a skill for independent practice. What constitutes a valid and reliable
302 assessment instrument is a well-established discussion in the behavioral sciences and has resulted in
303 international standards for testing²⁰. Application of these standards in medical education research and
304 practice has not been consistent.

305 As stated above, an assessment instrument should meet two requirements of inter-rater
306 reliability to be used in high-stakes assessments: 1) inter-rater reliability should be at least .90²¹ and 2)
307 this reliability should be based on the amount of agreement between the observers rather than the
308 amount of association between the scores²². Only 14% of the reported inter-rater reliability estimates
309 in our review were above .90 and based on agreement (including the ICC). Also, a substantial amount
310 of the documents lacked information necessary to summarize the information in a meta-analysis
311 statistically. This resulted in a marked reduction of the number of documents that could be included in
312 our meta-analysis: only 14 out of 247 documents.

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

322 **Limitations of evidence**

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

333 assessment situation (e.g., in vivo or simulation), the procedure that is performed, and the experience
334 level of participants and raters.

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

342 **Guidelines for the reporting of inter-rater reliability**

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

- 347 (1) Specify the subject population of interest: number of participants used for inter-rater reliability
348 evaluation, participants' level of experience, and demographics.
- 349 (2) Specify the assessor population of interest: number of assessors, assessors' level of experience,
350 and demographics.
- 351 (3) Describe the assessment process: blinding and training of assessors, how assessors were assigned
352 to participants (was the design fully crossed? See Hallgren ¹⁵).
- 353 (4) State the number of replicate observations.
- 354 (5) State which reliability index was used to evaluate inter-rater reliability. Report inter-rater
355 agreement rather than inter-rater consistency or association.
 - 356 a. Percentage or proportion agreement: report i) the estimate, ii) the sample size, and iii) the
357 number of observations per participant.
 - 358 b. Cohen's kappa: report i) the estimate, ii) the percentage or proportion agreement, iii) the
359 sample size, and iv) the number of observations per participant.

360 c. IntraClass Correlation (ICC): report i) the type of ICC according to the classification by
361 McGraw and Wong ³⁴, ii) the estimate, iii) the sample size, and iv) the number of
362 observations per participant.

363 (6) Provide information about the statistical precision of measurement. Report either a standard error
364 or a confidence interval.

365 **Strengths and limitations**

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

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

379 **Conclusion**

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

388

389

Acknowledgements

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

References

- 392
- 393 1. Giddings T, Gray G, Maran A, Mulligan P, Wheatley D, Williams JL. Response to the general
394 medical council determination on the Bristol case: Senate paper 5. *The Senate of Surgery of*
395 *Great Britain and Ireland*. 1998.
- 396 2. Kohn LT, Corrigan JM, Donaldson MS. *To err is human*. National Academies Press:
397 Washington, DC; 2000. doi:10.17226/9728
- 398 3. Institute of Medicine, Committee on the Governance and Financing of Graduate Medical
399 Education. *Graduate medical education that meets the nation's health needs*. Washington, DC:
400 Institute of Medicine; 2014.
- 401 4. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73.
402 doi: 10.1111/jedm.12000
- 403 5. Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective.
404 *BMJ*. 1999;318(7188):887-888. doi: 10.1136/bmj.318.7188.887
- 405 6. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective
406 assessment of technical surgical skills. *Br J Surg*. 2010;97(7):972-987. doi:10.1002/bjs.7115
- 407 7. Gallagher AG, Seymour NE, Jordan-Black J-A, Bunting BP, McGlade K, Satava RM.
408 Prospective, randomized assessment of Transfer of Training (ToT) and Transfer Effectiveness
409 Ratio (TER) of virtual reality simulation training for laparoscopic skill acquisition. *Ann Surg*.
410 2013;257(6):1025-1031. doi:10.1097/SLA.0b013e318284f658
- 411 8. MacMillan AIM, Cuschieri A. Assessment of innate ability and skills for endoscopic
412 manipulations by the advanced dundee endoscopic psychomotor tester: Predictive and
413 concurrent validity. *Am J Surg*. 1999;177(3):274-277. doi:10.1016/S0002-9610(99)00016-1
- 414 9. Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of
415 psychomotor skills in medical trainees: A systematic review. *Med Educ*. 2013;47(7):650-673.
416 doi:10.1111/medu.12220
- 417 10. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment
418 of procedural skills: A systematic review. *Am J Surg*. 2011;202(4):469-480.
419 doi:10.1016/j.amjsurg.2010.10.020

- 420 11. Reznick RK, Macrae H. Teaching surgical skill – Changes in the wind. *NEJM*. 2006;355:2664-
421 2669. doi: 10.1056/NEJMra054785
- 422 12. Swanson DB, van der Vleuten CPM. Assessment of clinical skills with standardized patients:
423 State of the art revisited. *Teach Learn Med*. 2013;25(sup1):S17-S25.
424 doi:10.1080/10401334.2013.842916
- 425 13. Streiner DL, Norman GR, Cairney J. *Health measurement scales: a practical guide to their*
426 *development and use* (3rd ed.). Oxford University Press: Oxford; 2003.
- 427 14. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and
428 reliable? *Injury*. 2011;42(3):236-240. doi:10.1016/j.injury.2010.11.042
- 429 15. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial.
430 *Tutor Quant Methods Psychol*. 2012;8(1):23. doi:10.20982/tqmp.08.1.p023
- 431 16. Kennedy A-M, Carroll S, Traynor O, Gallagher AG. Assessing surgical skill using bench
432 station models. *Plast Reconstr Surg*. 2008;121(5):1869-1870. doi:
433 10.1097/PRS.0b013e31816b19bc
- 434 17. Cicchetti D V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized
435 assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284-290. doi:10.1037/1040-
436 3590.6.4.284
- 437 18. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of
438 clinical measurement. *Lancet*. 1986;327(8476):307-310. doi: 10.1016/S0140-6736(86)90837-8
- 439 19. Gallagher AG, O’Sullivan GC, Leonard G, Bunting BP, Mcglade KJ. Objective structured
440 assessment of technical skills and checklist scales reliability compared for high stakes
441 assessments. *ANZ J Surg*. 2014;84(7-8):568-573. doi:10.1111/j.1445-2197.2012.06236.x
- 442 20. APA. NCME (American Educational Research Association, American Psychological
443 Association, & National Council on Measurement in Education). *Standards for educational*
444 *and psychological testing*. Amer Educational Research Assn; 1999.
- 445 21. Cooper C. *Individual differences and personality*. London: Arnold; 2002.
- 446 22. Altman DG. *Practical statistics for medical research*. Chapman & Hall/CRC: Boca Raton, FL;
447 1990.

- 448 23. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the
449 Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity
450 evidence. *Adv Heal Sci Educ.* 2015;20(5):1149-1175. doi:10.1007/s10459-015-9593-1
- 451 24. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement
452 Studies (GRRAS) were proposed. *Int J Nurs Stud.* 2011;48(6):661-671.
453 doi:10.1016/j.ijnurstu.2011.01.016
- 454 25. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical
455 residents' clinical competence. *Ann Intern Med.* 2004;140(11):874-881. doi:10.7326/0003-
456 4819-140-11-200406010-00008
- 457 26. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: Mechanisms that
458 contribute to assessor differences in directly-observed performance assessments. *Adv Heal Sci*
459 *Educ.* 2013;18(3):325-341. doi:10.1007/s10459-012-9372-1
- 460 27. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on
461 reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *J Gen Intern Med.*
462 2009;24(1):74-79. doi:10.1007/s11606-008-0842-3
- 463 28. VanBlaricom AL, Goff BA, Chinn M, Icasiano MM, Nielsen P, Mandel L. A new curriculum
464 for hysteroscopy training as demonstrated by an objective structured assessment of technical
465 skills (OSATS). *Am J Obstet Gynecol.* 2005;193(5):1856-1865. doi:10.1016/j.ajog.2005.07.057
- 466 29. Fialkow M, Mandel L, VanBlaricom A, Chinn M, Lentz G, Goff B. A curriculum for Burch
467 colposuspension and diagnostic cystoscopy evaluated by an objective structured assessment of
468 technical skills. *Am J Obstet Gynecol.* 2007;197(5):544.e1-544.e6.
469 doi:10.1016/j.ajog.2007.07.027
- 470 30. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than
471 hypothesis testing. *BMJ (Clin Res Ed).* 1986;292(6522):746-750.
472 doi:10.1136/bmj.292.6522.746
- 473 31. Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and
474 appropriately quantified. *Int J Epidemiol.* 2008;37(5):1158-1160. doi: 10.1093/ije/dyn204
- 475 32. Sun S. Meta-analysis of Cohen's kappa. *Heal Serv Outcomes Res Methodol.* 2011;11(3-4):145-

- 476 163. doi:10.1007/s10742-011-0077-3
- 477 33. Shrouf PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.*
478 1979;86(2):420. doi:10.1037/0033-2909.86.2.420
- 479 34. McGraw KO, Wong SP. Forming inferences about some intraclass correlations coefficients.
480 *Psychol Methods.* 1996;1(1):30-46. doi:10.1037/1082-989X.1.4.390
- 481 35. Viechtbauer W. Conducting Meta-Analyses in R with the **metafor** Package. *J Stat Softw.*
482 2010;36(3). doi:10.18637/jss.v036.i03
- 483 36. Seymour NE, Gallagher AG, Roman S a, et al. Virtual reality training improves operating room
484 performance: results of a randomized, double-blinded study. *Ann Surg.* 2002;236(4):458-63;
485 discussion 463-4. doi:10.1097/01.SLA.0000028969.51489.B4
- 486 37. Sarker SK, Chang A, Vincent C, Darzi SAW. Development of assessing generic and specific
487 technical skills in laparoscopic surgery. *Am J Surg.* 2006;191(2):238-244.
488 doi:10.1016/j.amjsurg.2005.07.031
- 489 38. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training
490 significantly reduces the error rate for residents during their first 10 laparoscopic
491 cholecystectomies. *Am J Surg.* 2007;193(6):797-804. doi:10.1016/j.amjsurg.2006.06.050
- 492 39. Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in
493 mastoidectomy. *Laryngoscope.* 2009;119(12):2402-2410. doi:10.1002/lary.20678
- 494 40. Andersen SAW, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of
495 virtual simulation training using final-product analysis. *Laryngoscope.* 2015;125(2):431-435.
496 doi:10.1002/lary.24838
- 497 41. Wong DM, Watson MJ, Kluger R, et al. Evaluation of a task-specific checklist and global
498 rating scale for ultrasound-guided regional anesthesia. *Reg Anesth Pain Med.* 2014;39(5):399-
499 408. doi:10.1097/AAP.000000000000126
- 500 42. Angelo RL, Pedowitz RA, Ryu RKN, Gallagher AG. The Bankart performance metrics
501 combined with a shoulder model simulator create a precise and accurate training tool for
502 measuring surgeon skill. *Arthroscopy.* 2015;31(9):1639-1654. doi:10.1016/j.arthro.2015.04.092
- 503 43. Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG. The Bankart performance metrics

- 504 combined with a cadaveric shoulder create a precise and accurate assessment tool for
505 measuring surgeon skill. *Arthroscopy*. 2015;31(9):1655-1670. doi:10.1016/j.arthro.2015.05.006
- 506 44. Day RW, Fleming J, Katz MH, et al. Rapid assessment of technical competency: The 8-min
507 suture test. *J Surg Res*. 2015;200(1):46-52. doi:10.1016/j.jss.2015.06.057
- 508 45. Fried MP, Sadoughi B, Gibber MJ, et al. From virtual reality to the operating room: The
509 endoscopic sinus surgery simulator experiment. *Otolaryngol - Head Neck Surg*.
510 2010;142(2):202-207. doi:10.1016/j.otohns.2009.11.023
- 511 46. Iordache F, Bucobo JC, Devlin D, You K, Bergamaschi R. Simulated training in colonoscopic
512 stenting of colonic strictures: Validation of a cadaver model. *Color Dis*. 2015;17(7):627-634.
513 doi:10.1111/codi.12887
- 514 47. Ma IWY, Zalunardo N, Pachev G, et al. Comparing the use of global rating scale with
515 checklists for the assessment of central venous catheterization skills using simulation. *Adv Heal*
516 *Sci Educ*. 2012;17(4):457-470. doi:10.1007/s10459-011-9322-3
- 517 48. Koehler RJ, Nicandri GT. Using the arthroscopic surgery skill evaluation tool as a pass-fail
518 examination. *J Bone Joint Surg Am*. 2013;95(23):e187.
- 519
- 520

521

Supplementary Material

522 SM1 Search strategy

523 SM2 Inter-coder agreement for the general information categories and the primary and secondary

524 outcome categories

525 SM3 PRISMA 2009 Checklist.

526 SM4 Table. Number of documents reporting overall reliability estimates for inter-rater reliability by

527 reliability index and assessment instrument (N = 247)

528

529 **Figure legends**

530 Figure 1. PRISMA flow diagram for the selection of documents.

531 Figure 2. Distribution of the 225 agreement based reliability estimates.

532 Figure 3. PRISMA flow diagram for the selection of documents for the meta-analyses.

533

534 Table 1. Study characteristics of the studies included in the meta-analyses.

Study	Year	Assessment	Reliability	Assessment	Participants	Sample			
		instrument	index	situation		size	Assessors	Training	Blinding
Procedure-specific checklists									
Seymour NE ³⁶	2002	Task-specific checklist	Proportion agreement	In vivo/image-guided	Residents	16	Consultants	Yes	Yes
Sarker SK ³⁷	2005	Task-specific checklist	Cohen's kappa	In vivo/image-guided	Consultants	8	Consultants	Unknown	Yes
Ahlberg G ³⁸	2007	Task-specific checklist	Proportion agreement	Simulated/image-guided	Residents	13	Experts	Unknown	Yes
Laeq K ³⁹	2009	Task-specific checklist	Proportion agreement	Simulated/open	Residents	23	Unknown	Unknown	Yes
Gallagher AG ¹⁹	2014	Task-specific checklist	Proportion agreement	Simulated/open	Residents	19	Consultant	Yes	No
Andersen SA ⁴⁰	2015	Task-specific checklist	Cohen's kappa	Simulated/open	Residents	34	Experts	Unknown	Yes

Wong IH ⁴¹	2014	Task-specific checklist	Cohen's kappa	Simulated/image-guided	Medical students	35	Consultant	Unknown	Yes
Angelo RL ⁴²	2015	Task-specific checklist	Proportion agreement	Simulated/image-guided	Mixed	19	Consultant	Yes	Yes
Angelo RL ⁴³	2015	Task-specific checklist	Proportion agreement	Simulated/image-guided	Mixed	22	Consultant	Yes	Yes
Day RW ⁴⁴	2016	Task-specific checklist	Cohen's kappa	Simulated/open	Mixed	41	Mixed	Unknown	Yes

Rating scales

Laeq K ³⁹	2009	Global rating scale	Proportion agreement	Simulated/open	Residents	23	Unknown	Unknown	Yes
Fried MP ⁴⁵	2010	Global rating scale	Cohen's kappa	Combination	Residents	25	Experts	Unknown	Yes
Gallagher AG ¹⁹	2014	OSATS global rating scale	Proportion agreement	Simulated/open	Unknown	19	Unknown	Yes	No
Wong IH ⁴¹	2015	Global rating scale	Cohen's kappa	Simulated/image-guided	Medical students	35	Consultant	Unknown	Yes

Iordache F ⁴⁶	2015	Task-specific rating scale	Cohen's kappa	Simulated/image- guided	Mixed	20	Other	Unknown	Unknown
Other instruments									
Laeq K ³⁹	2009	Pass/fail decision	Proportion agreement	Simulated/open	Residents	23	Unknown	Unknown	Yes
		Pass/fail decision	Proportion agreement	Simulated/open	Residents	23	Unknown	Unknown	Yes
Ma IW ⁴⁷	2012	Pass/fail decision	Cohen's kappa	Simulated/open	Residents	34	Consultants	Yes	Yes
Koehler RJ ⁴⁸	2013	Pass/fail decision	Proportion agreement	Simulated/image- guided	Mixed	30	Unknown	Unknown	Yes
Wong IH ⁴¹	2015	Pass/fail decision	Cohen's kappa	Simulated/image- guided	Medical students	35	Consultant	Unknown	Yes

535

536

537

538 Table 2. Pooled inter-rater reliability estimates and confidence intervals (CI) for multilevel random
 539 effects regression models for Cohen's kappa and proportion agreement.

Model	n	Pooled estimate	CI	<i>p</i>-value
Cohen's kappa				
No moderators	7	.78	.69 - .89	< .001
Moderator: checklists	4	.82	.69 - .95	< .001
Moderator: rating scales	3	.79	.63 - .95	< .001
Moderator: other instruments	2	.61	.37 - .86	< .001
Proportion agreement				
No moderators	6	.84	.71 - .96	< .001
Moderator: checklists	5	.84	.72 - .97	< .001
Moderator: rating scales	2	.69	.52 - .86	< .001
Moderator: other instruments	2	1.0	.84 - 1.2	< .001

540