Education in laparoscopic surgery

Kramp, Kelvin Harvey

Chapter 6

ESTIMATING THE INTER-RATER RELIABILITY OF SURGICAL SKILLS ASSESSMENT

Kelvin H. Kramp, Marc J. van Det, Jean-Pierre E.N. Pierie

Submitted

Abstract

The interest in the reliability of surgical skills assessment has increased substantially over the past decades. Inter-rater reliability, a subform of reliability, is defined as the amount of agreement between human raters using the same assessment instrument. We discuss important aspects of the statistics and study design in the context of subjective assessment in surgical education. The aim of this paper is to equip the surgeon scientist with the statistical methods and study designs for evaluating the inter-rater reliability of surgical skills assessment and to provide designers of surgical training programs and clinical supervisors with the necessary knowledge for assessing the quality of these studies.

1. Background

The majority of current surgeons were trained according to the master-apprentice model in which a master surgeon decides whether a trainee showed sufficient improvement based on his/her own perception of the necessary skills and knowledge for surgery. However, pressure from the public and governmental institutions has led to the development of more objective assessment methods in the last decennia. [1,2] These surgical assessment methods force clinical supervisors to quantify the quality of the observed skills on a specific set of domains relevant to the development of surgical competency. The numerical ratings can be used to monitor progression during a training program, identify strengths and weaknesses in trainees, compare the efficacy of different training curriculums, measure retention of skills after a training program and facilitate licensing in the independent treatment of uncomplicated disease.

Multiple studies have shown that a proportion of these methods are valid tools for some of these purposes. [3–8] Unfortunately, a concern repeatedly addressed in these studies is the insufficient amount of agreement between raters rating the same performance, a concept also known as inter-rater reliability.[1] Inadequate reliability can impede implementation of an assessment method, because the outcome of an assessment can only be utilized if the precision is of an acceptable level. While the introduction of simulators created an opportunity for more objective and reliable assessment of psychomotor skills, the assessment of higher levels of cognitive abilities still remains a task of experienced surgeons charged with the responsibility to safely guide trainees during the acquisition of surgical skills in the highly dynamic environment of the OR. New assessment methods are continually being developed and improved to increase the inter-rater reliability of assessment by supervising surgeons. However, although reliability coefficients, which are used to measure inter-rater reliability, are one of the most important aspects of assessment methodology, those involved in surgical education are frequently unfamiliar with the rationale behind the statistics and study designs necessary for the execution of reliability research of an acceptable quality. Also, the surgeon scientist, eager to conduct research according to the highest scientific principles, can be confronted with the difficulties of choosing the right combination of statistics and methodology to estimate the validity and inter-rater reliability for a study focused on subjective assessment methodology. This can pose a problem in the field of surgical education, because understanding of the calculation methods of the intra-class correlation coefficient (ICC), the most often used statistic for calculating inter-rater reliability, is imperative for correct execution and interpretation of studies addressing the inter-rater reliability of surgical skills assessment. Previous reviews published in the surgical literature that addressed inter-rater reliability were limited in promoting a deeper understanding of inter-rater reliability. [9,10] Therefore, the aims of this paper are:

1) To provide an introduction to the use and rationale of the ICC.

2) To discuss important aspects of study design in the calculation and evaluation of inter-rater reliability.

2. The ICC

2.1 Rationale behind the ICC

Of 52 studies included in two systematic reviews addressing inter-rater reliability of subjective assessment in surgical education, 22 studies used the ICC, 13 studies used Cronbach alpha, 3 studies used a Pearson or Spearman correlation coefficient, 3 studies used Generizability coefficient and 11 studies used various other methods. [1,2] The ICC can therefore be considered as the most frequently used measure of inter-rater reliability in surgical education.

There are advantages of choosing the ICC to calculate inter-rater reliability instead of other correlation coefficients. The disadvantage of using the Pearson correlation coefficient for inter-rater reliability is that it only estimates the degree of association between two variables and says little about the amount of agreement between measurements. The Pearson correlation coefficient, however, continues to be used by some as a measure of inter-rater reliability until recently [27–30], while it can better be reserved for the quantification of an association between two measurements that do not share metric or variance (e.g. BMI and daily caloric intake). To estimate the correlation between measurements that have the same unity, or belong to the same 'class', such as measurements performed by different raters on the same scale, the ICC is a better candidate. [3] Moreover, the ICC also has the advantage of being able to measure the correlation between more than two series of measurements (e.g. ratings from 3 or more raters), while the Pearson and Spearman correlation coefficients are limited to the use for two variables.

Generizability theory is an upcoming theory for estimating reliability in the field of educational psychology. Generizability theory is based on the estimation of variance components. It gives researchers the opportunity to calculate the exact percentage of error variance each source of error is responsible for. Although these opportunities make generizability theory an attractive model, the calculation method used for estimation of variance components used in generazibility varies between and within software packages (e.g. ANOVA, maximum likelihood and minimum norm quadratic unbiased estimation) while the ICC models are only based on ANOVA calculations. Moreover, some of the software packages such as SPSS and SAS are unable to cope with missing data when calculating variance components, which is a relatively common phenomenon in educational research. And third, calculation of the reliability of the assessment of a single rater (similar as the single measures ICC models), which is the measure of interest in the majority of validation studies, requires the execution of additional so-called Decision-studies, while the ICC calculations with standard software packages directly provide estimates for the use of single and average ratings.

The numerical value of the ICC can be calculated with different models. To get a general idea of what is measured with these models, the reliability coefficient calculated with the ICC can be simplified to:

$$\textit{Reliability coefficient} = \textit{True variance} / [\textit{True variance} + \textit{Error variance}] \quad 1)$$

Thus, the reliability coefficient calculated with the ICC is in essence the proportion of variance in the sample attributable to true variance. True variance is an abstract concept, but it can be estimated by subtracting the error variance from the total variance. The formula for the reliability coefficient would then become:

$$\textit{Reliability coefficient} = [\textit{Total variance} - \textit{Error variance}] / [\textit{Total variance}] \quad 2)$$

The resulting reliability coefficient is a number between 0 and 1, whereby 0 means no agreement between measurements and 1 means total agreement between measurements. The exact formulas for calculating the ICC can be found in the publication of Shrout&Fleiss. [11]

2.2 How to choose the right model to calculate the ICC

In 2 systematic reviews addressing the validity and reliability of surgical assessment, 17 out of the 22 studies that used the ICC to calculate a reliability coefficient did not report the used calculation model. [1,2] However, the inter-rater reliability can vary significantly depending on the model used to calculate the reliability coefficient. Lahey et al. have reported examples of 20-fold differences in size while using the same data. [12] It is therefore important to choose the right model according to the design of the study and to report which model has been used so the appropriateness of the applied ICC model can be evaluated as a part of quality assessment.

In total there are 6 different formulas to calculate the ICC: ICC-1 to ICC-3, which have different assumptions concerning raters and subjects, and type 1 or type k, which indicate a single or average measures ICC. A flowchart for choosing a model and examples of application of these models in the research field of surgical education are provided in figure 1.

Model 1 (one-way random) is suitable when the same subjects are rated by different raters during the study. This calculationmodel assumes subjects are not consistently rated by the same set of raters in the research setting. The calculationmodel allows generalization to other raters and subjects with the same characteristics, but it does not enable the users of the assessment instrument to mathematically infer the specific part of error variance that can be attributed to the variance between raters in the resulting reliability coefficient. Model 2 (two-way random), referred to as absolute agreement model, can be applied when the same subjects are all rated by the same set of raters during the study. For model 3 (two-way mixed), referred to as consistency agreement model, the same is true as for model 2, except that in this model the raters are assumed to be the exact same raters that will conduct assessment in the future. This model can only be used in the case that the included raters will be the only raters that will perform assessments in the future or the researcher is not interested in the absolute differences between ratings, but only in the inconsistencies between ratings.

The 3 models can be used to calculate the reliability for 2 types of measurements: single measures (type 1) and average measures (type k). Mathematically the type 1 coefficients are a derivative of type k coefficients. The average measures ICC will always give a higher estimate then the single measures ICC, as average measures are more reliable than single measures, however, the average measures ICC can only be used in special cases. If a researcher has used the average of a series of measurements to calculate a mean outcome to estimate reliability, and secondly, will also apply the same protocol in the future, the average measures ICC is of interest. If one of these criteria is not met, the single measures ICC is of interest. Interestingly, the average measures ICC of model 3 (ICC-3,k) is mathematically equal to Cronbach alpha. [13]

It is important to note that, that the ‘subjects’ do not per se have to consist of different persons. For instance, the ratings of one subject rated during different levels of experience can also be treated as multiple subjects in the calculation model.

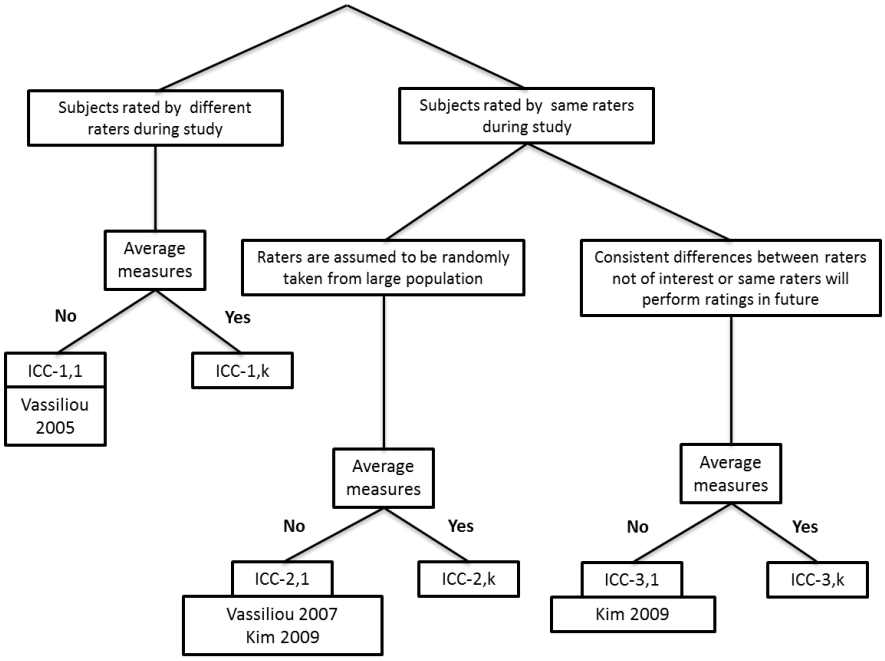

Figure 1: Flowchart for choosing a model based on study design and examples from the literature.

2.3 Calculating a sample size

Clinical supervisors can only invest a limited amount of time because of clinical burdens and there are typically only a limited number of consultant surgeons available or willing to participate. It is therefore likely that, as in most studies in medical research, the smaller the sample size the higher the feasibility of the study will be. To ensure that the sample of subjects to be rated is sufficient to achieve statistical significant results, a sample size calculation can be conducted with complex calculations published by Donner&Eliasziw or in non-integral values with the more simple, but less exact, formula published by Walter et al. [14,15]:

$$k = 1 + (2\,(2.4865)^2\,N / (\ln C_0)^2\,(n - 1)) \qquad 3)$$

where, k = number of needed subjects at $\alpha = 0.05$ (significance level, type I error or false positive rate) and $\beta = 0.20$ (power, type II error or false negative rate), N = number of raters and C_0 is given by:

$$C_0 = (1+ (N\,(ICC_0 / (1 - ICC_0)))) / (1+ (N\,(ICC_1 / (1 - ICC_1)))) \qquad 4)$$

where, ICC_0 = ICC of the null hypothesis (H_0) and ICC_1 = ICC of the alternative hypothesis (H_1). In the case of N = 2 there are small adjustments in the formula for the sample size calculation. [14] Table 1 shows the calculated sample sizes for different values of N for H_0:ICC = 0 and for H_1:ICC = 0.4 or 0.8. Note that, as is true for correlations in general, smaller differences between H_0 and H_1 require larger sample sizes and that small sample sizes require higher ICCs to reach statistical significance.

The sample size calculation with the formula of Walter et al. is based on ICC-1. Because this model results in the smallest ICC, it can also be used as a practical method to estimate the sample size for ICC-2 and 3. Furthermore, sample sizes are the same for the single and average measures type models.

Table 2: Sample sizes calculated with formula 3 and rounded to integral values for $ICC_0 = 0$ and ICC_1 values of 0.4 and 0.8 at $\alpha = 0.05$ and $\beta = 0.20$. [14]

Numer of raters	Sample size	
	ICC_1	
	0.4	0.8
3	16	4
4	11	3
5	8	3
10	4	2

2.4 Interpretation of the size of the ICC

The reliability coefficient indicates the proportion of the variance that can be attributed to true variance. The remaining proportion of variance can be caused by rater error, random error and/or other sources of error. If the reliability coefficient is very high or low, it is less difficult to draw conclusions than when the reliability coefficient is in between extreme values. Cut-off values used for classification of the reliability coefficient can be helpful, but are always arbitrary in nature and should be adjusted to the purpose of the measurement instrument. For formative assessment (feedback during learning), the interpretation values may be less stringent than for summative assessment (high stakes examination). [16] In surgical and medical literature, a cut-off value of 0.8 has widely been adopted as the threshold for high stakes examination [1,9,17–20], although there is no high level evidence that supports a rationale for this specific value. [21]

Another option for interpreting the reliability coefficient, is to calculate the standard error of measurement (SEM). The SEM can be used to assess the corresponding probability distribution of the obtained score of a subject in the case that consecutive assessments would be conducted on the same subject by other raters (assuming these raters have similar characteristics as those included in the original research). The SEM can be calculated with the formula [22]:

$$SEM = Sd \times \sqrt{(1-ICC)} \qquad 5)$$

Where Sd = the standard deviation of scores calculated for a set of ratings on the performance level of the subject of interest. This method allows a more exact interpretation of results. The SEM can be used to assess the 95% confidence interval of a single rating of a subject, assuming the rater and subject have the same characteristics as those that participated in the reliability study. Let's take the example of an assessment score with the OSATS of 11/35 of a novice (35 is the maximum score of the OSATS). If in previous studies it has been shown that the Sd for novices is 3 and the inter-rater reliability of the OSATS is 0.58, the SEM would be 1.94 and the corresponding 95% confidence interval for the assessment repeated by other raters would be 7/35 – 15/35 in rounded scores.

2.5 P-values of the ICC

Just as the p-values of the t-test are based on a t-value, the p-values of ICC-1, -2 and -3 are based on the F-value of the ANOVA models (resp. one-way random, two-way random or two-way mixed ANOVA). The F-value is calculated with the mean variance components described in table 2. If the p-value of the F-test is not significant at the corresponding degrees of freedom, which is based on the number of subjects and raters, there could be insufficient variance between subjects to calculate a reliability coefficient and the coefficient should be looked at with skepticism. Reliability is defined by the amount of agreement between ratings, but is also dependent on the true variance within the sample. True variance in assessment scores can be jeopardized as a consequence of the tendency of

assessors to rate all trainees as average during a live observation. This is also known as the 'central tendency error' and has been reported as a problem in in-training evaluation reports (ITER) by some authors. [23–25] Participants should therefore be stimulated to use the full range of the scales as much as possible and psychosocial barriers for rating trainees as below or above average should be evaluated and managed appropriately.

In ICC-2 and -3, it can additionally be useful to look at the p-value of the F-test for the variance between raters. Opposite to the F-test for the total variance, this p-value should not be significant to indicate there is no significant difference between the assessment scores of raters.

Table 2: Six different ICCs: 3 different models (ICC-1 to ICC-3), all of 2 different types (type 1 and type k). Var = Mean Square (Mean variance). ANOVA = ANalysis Of VAriance. S = Size

ICC	ANOVA	Raters	Subjects	Agreement	Total variance		Error component		S
					ANOVA	*ICC*	*ANOVA*	*ICC*	
1-1	One-way random		Random		Between-groups var	Between- subjects var	Within-group error	Within-subjects error	↓
1-k	One-way random		Random		Between-groups var	Between- subjects var	Within-group error	Within-subjects error	
2-1	Two-way random	Random	Random	Absolute agreement model	Within-subjects var	Between-subjects var	Between-subjects var & Within-subjects error	Between raters var & residual error	
2-k	Two-way random	Random	Random	Absolute agreement model	Within-subjects var	Between-subjects var	Between-subjects var & Within-subjects error	Between raters var & residual error	
3-1	Two-way mixed	Fixed	Random	Consistency agreement model	Within-subjects var	Between-subjects var	Within-subjects error	Error	
3-k	Two-way mixed	Fixed	Random	Consistency agreement model	Within-subjects var	Between-subjects var	Within-subjects error	Error	

2.6 Evaluation of factors influencing the ICC

When two or more assessment methods used by a sample of raters to rate a sample of subjects, seem to differ in terms of reliability, it is sensible to check whether the difference does not originate from a difference in true variance by evaluating the total variance of ratings of the two assessment methods. In figure 2, an example is shown of a hypothetical study in which 3 assessment forms are used: 1) a procedural-based assessment (PBA) for appendectomy, 2) a PBA for hemicolectomy and 3) a global ratings scale (GRS). The scores shown are based on the mean scores of multiple raters. If we assume that the amount of rater error and random error are equal for the appendectomy PBA and the GRS, the reliability of the former would automatically be higher as a consequence of the larger 'true' variance in scores (0-95% for the PBA vs. 0-75% for the GRS). For the same reason the PBA for the appendectomy would be more reliable in the earlier stage and the PBA for the hemicolectomy more reliable in the later stage of surgical training.

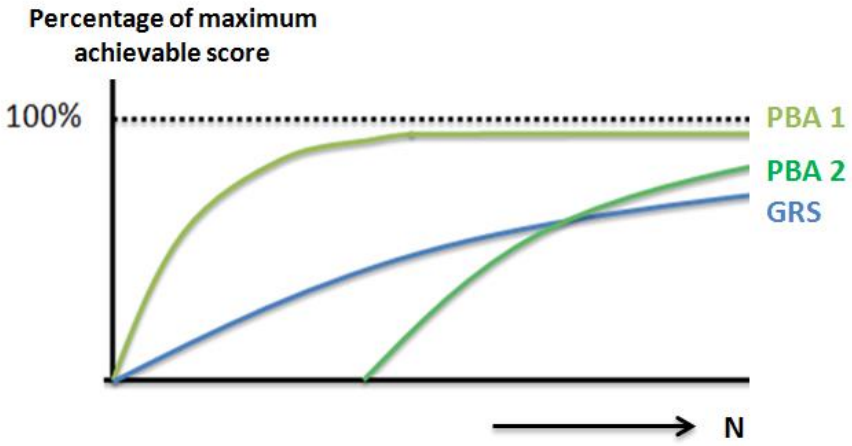

Figure 2: On the x-axis the experience of a trainee and on the y-axis the mean normalized performance level of 2 raters on a procedural-based assessment of an appendectomy (PBA 1), a procedural-based assessment of a hemicolectomy (PBA 2) and a global rating scale (GRS).

Besides true variance, factors like study design, rater background or rater setting can also significantly influence the reliability coefficient in the wrong or right direction. Whether one or more factors had a dominant effect on the outcome can be evaluated with a correlation matrix (Table 3). However, given k raters a correlation matrix consists of (k-1)*k/2 ICCs. In the case of a very large quantity of raters it can therefore be valuable to create a geometrical representation by plotting the ICCs between raters as vectors in a graph (Figure 3). [26] To achieve a geometrical rendition, the correlations between raters can be calculated into degrees by using the inverse cosine function ($\cos^{-1}$). The smaller the correlation between variables the larger the angle between the vectors in the graphical representation will be.

Table 3: A correlation matrix of ICC-2,1 of 6 raters performing 20 consecutive assessments. V = Video rater, D = Direct observer. Time point functions as an interaction effect and reduced inter-rater reliability in the group of video raters during the last 5 to 10 assessments (V1-V3).

Raters	Time point			
	1-5	***6-10***	***11-15***	***16-20***
V1,V2,V3	0.90	0.85	0.50	0.12
D1,D2,D3	0.95	0.91	0.89	0.97

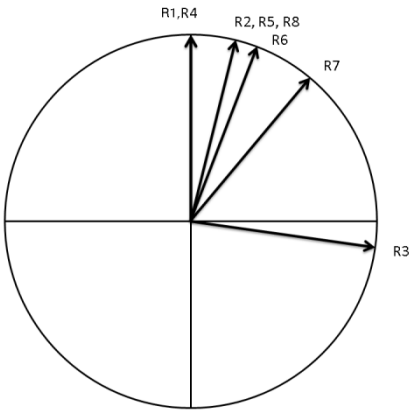

Figure 3: Geographical representation of correlations between 8 raters. R = Rater. The correlation between raters is translated into degrees with the inverse cosine function ($\cos^{-1}$) to obtain vectors.

3. Study design

There are a number of factors that can result in bias or concerns of applicability of the reliability coefficient reported by studies that address the assessment of surgical skills (Table 4).

Table 4: Issues pertinent to the evaluation of study quality.

Risk of bias	
Study design	Blinding: - Subject identity - Handedness - Skin color - Time duration of performance - Voice
	Randomization: - Sequence of subject performances - Sequence of assessment forms - Raters
Statistics	If ICC used as measure of inter-rater reliability: chosen ICC model reported?
Concerns of applicability	
Study design	Participants: - Rater characteristics - Subject characteristics - Motivation of raters
	Training: - Training content described in sufficient detail - Instructions to use full range of scores
Statistics	If ICC used as measure of inter-rater reliability: Appropriate ICC model used

3.1 Transparency

Because there are so many different modalities for calculating reliability, transparency is essential for the assessment of study quality. It can be tempting to use the wrong ICC model, as some models tend to give higher results than others. When Generalizability theory is used, important aspects to report are whether facets are crossed or nested, whether facets are fixed or random, the existence of negative variance components and which software was used. Studies using a Pearson or Spearman correlation coefficient not reporting which of the 6 mathematical different ICCs was used, or not describing the methodological route for estimating generazilibilty coefficients should be classified as a study of low quality or at risk of bias. [31,32]

3.2 Randomization and blinding

An assessment method can be evaluated with the raters blinded or not blinded to the identity or experience level of the trainee. A meta-analysis of randomized clinical trials that compared blinded versus non-blinded observer assessments of subjective measurement scales showed that the effect size was exaggerated with 68% in the group of non-blinded studies. [33] Although the number of studies in the surgical literature investigating the effect of concealment of identity is limited, it is reasonable to assume that knowing the identity of the trainee can influence the outcome of assessment. The assessment can theoretically be biased in the positive (Halo effect) or in the negative direction (Devil effect) due to awareness of the identity of the trainee. The risk of introducing bias in the assessment of surgical skills can be avoided in the assessment of surgical skills by using video images restricted to the hands without revealing skin color or handedness, as demonstrated by Vogt et al. [34] In the case of laparoscopic surgery, the laparoscope can be used to record a video restricted to images of the inside of the abdomen. Using blinded videos can pose a

problem when items of an assessment incorporate elements of communication of the trainee with the operating team. If verbal elements of communication are to be assessed, a sound recording of the communication can be used to subtitle the video or video parts involving communication can be tagged.

A remaining potential source of bias in the assessment of blinded videos is the time length of the video. Based on the duration of a video, one can estimate roughly the experience level, assuming the difficulty of the task is equal. This can partially be circumvented by using video fragments according to a well-defined protocol. However, editing videos can in turn threaten generalization to the assessment of a whole procedure.

Other potential sources of bias in blinded assessments include the sequence of the performances to be assessed and the sequence of the assessment forms. In drug innovation, there is high level evidence of an exaggeration of the effect size in studies with unclear or inadequate random sequence generation. [35] To avoid raters using the sequence of the performances as a source of information for assessment, performances can be randomized. In the case that multiple assessment forms are used simultaneously, raters can develop a raised subconscious or conscious awareness of the strengths and weaknesses during the completion of the first assessment that is transferred to the subsequent assessment. To minimalize the chance that the order of the videos or the assessment methods influences reliability, the sequence of assessment can be randomized. An elaborate description of different randomization methods and guidelines on which randomization method to choose according to the study design has been published by Kao et al. [36]

3.3 Participant characteristics

To avoid concerns of applicability, the included subjects and raters should have characteristics similar to the population of interest. When subjects are chosen, the range of experience levels of subjects should be similar to the range of experience levels in the population in which the measurement instrument is going to be used. For example, demonstrating that an assessment method can reliable assess an expert performance is futile if the assessment method is designed for tracking improvement during training.

A number of authors in surgical education and applied psychology have suggested on the basis of their findings that as raters become more familiar with the assessment method, the reliability of the assessment increases. [8,38–42] In a recent study in cardiothoracic surgery that investigated the influence of rater training on the reliability of assessment, a dramatic increase in reliability coefficients was observed after training from 0.09-0.48 to 0.80-0.90. [20] Therefore, when a new assessment form is tested, or an already validated assessment form is evaluated in another population of raters, training of raters can be necessary to obtain maximum accuracy of assessment scores. Whenever training is relevant for the accuracy of assessment, it is important to report the content of rater training in sufficient detail to allow replication of training in other settings.

Fatigue or a lack of motivation can endanger the accuracy of assessment inside and outside the research context. An assessment form should therefore be able to be completed within a feasible time frame. On the other hand, the internal consistency, measured with Cronbach alpha, is an important aspect of reliability other than inter-rater reliability and tends to increase with the number of items that measure the same trait. The internal consistency rises because measurement errors of the individual items will tend to cancel each other out as the number of items that measure the same trait increases, leading to a more accurate estimate of the measured trait. [26] Choosing the total number of items therefore includes finding the optimum balance between feasibility and reliability.

The use of extrinsic rewards can be a valuable instrument to increase interest in participation among surgeons when there is an initial lack of motivation or to optimize commitment and persistence during assessment. [43] Care should be taken to avoid diminishing initial intrinsic motivation among volunteering research participants by introducing an (inadequate) external reward, a phenomenon that has been observed in the field of cognitive psychology, and has become known as the ‘overjustification effect’. [44–46]

3.4 Constructivist social-psychological approach

The points raised in this review have primarily been described from a psychometric perspective. Although raters are seen as measurement instruments in the psychometric approach, raters have unique cognitive processes during assessment. [47] These cognitive processes of assessors are influenced by the acquired knowledge during training within one or more educational institutions, personal operative experiences during their surgical career and the content and characteristics of the interactions with surgical supervisors who trained and supervised them in the skillslab and/or in the OR (socialization). Goovaerts et al. justifiably stated that, although the ratings do not agree from a psychometric standpoint of view, they can all be equally valid and might separately all contribute to a more complete picture of the quality of surgical skills. In our endeavours to objectify surgical skills, we should not forget that the art of surgery can never fully be expressed in something as simple as a number.

4. Conclusions

The public and government have acknowledged that training is a vital aspect of effective patient care and have therefore urged for more objective quality assessment during surgical education. As a consequence, research in surgical education has provided, and will continue to provide, tools to quantify the quality of surgical skills. This review recapitulates on the statistics and study design behind the inter-rater reliability from educational measurement and describes important aspects of the statistics and study design of studies estimating the inter-rater reliability of surgical skills assessment. This paper is aimed to equip surgeon scientists with methods for investigating the inter-rater reliability of subjective assessment and provide designers of surgical training programs and clinical supervisors with the necessary skills for assessing the quality of these studies.

References

1. Hove, P. D. Van, Tuijthof, G. J. M., Verdaasdonk, E. G. G., Stassen, L. P. S. & Dankelman, J. Objective assessment of technical surgical skills. Br J Surg. 972–987 (2010).
2. Jelovsek, J. E., Kow, N. & Diwadkar, G. B. Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med. Educ.* 47, 650–673 (2013).
3. Martin, J. A. *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* 84, 273–278 (1997).
4. Niitsu, H. *et al.* Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today.* 43, 271–275 (2012).
5. Hopmans, C. J. *et al.* Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study. *Surgery.* 156, 1078-1088 (2014).
6. Hiemstra, E., Kolkman, W., Wolterbeek, R., Trimbos, B. & Jansen, F. W. Value of an objective assessment tool in the operating room. *Can J Surg.* 54, 116-122 (2011).
7. Vassiliou, M. C. *et al.* A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 190, 107-113 (2005).
8. Kramp, K. H. *et al.* Validity and Reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in Novice Trainees Performing a Laparoscopic Cholecystectomy. *J Surg Educ.* 72, 351-358 (2015).
9. Gallagher, A. G., Ritter, E. M. & Satava, R. M. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc. Other Interv Tech.* 17, 1525–1529 (2003).
10. Karanicolas, P. J. *et al.* Evaluating agreement: conducting a reliability study. *J Bone Joint Surg Am.* 91 Suppl 3, 99–106 (2009).
11. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull.* 86, 420 (1979).
12. Lahey, M. A., Downey, R. G. & Saal, F. E. Intraclass correlations: There's more there than meets the eye. *Psychol Bull.* 93, 586 (1983).
13. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1, 30 (1996).
14. Walter, S. D., Eliasziw, M. & Donner, a. Sample size and optimal designs for reliability studies. *Stat Med.* 17, 101–110 (1998).
15. Donner, A. & Eliasziw, M. Sample size requirements for reliability studies. *Stat Med.* 6, 441–448 (1987).
16. Downing, S. M. Reliability: on the reproducibility of assessment data. *Med Educ.* 38, 1006–1012 (2004).
17. Delfino, A. E., Chandratilake, M., Altermatt, F. R. & Echevarria, G. Validation and piloting of direct observation of practical skills tool to assess intubation in the Chilean context. *Med. Teach.* 35, 231-236 (2013).
18. Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D. & Ziv, A. Reliability estimates: behavioural stations and questionnaires in medical school admissions. *Med Educ.* 46, 277–288 (2012).
19. Arain, N. A. *et al.* Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surg Endosc.* 26, 2740–2745 (2012).
20. Lou, X. *et al.* Training less-experienced faculty improves reliability of skills assessment in cardiac surgery. *J Thorac Cardiovasc Surg.* 148, 2491-2496 (2014).
21. Norcini, J. J. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 74, 1088–1090 (1999).
22. Harvill, L. M. Standard Error of Measurement. *Educ Meas Issues Pract*. 33–41 (1991).

23. Feldman, L. S., Hagarty, S. E., Ghitulescu, G., Stanbridge, D. & Fried, G. M. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *J Am Coll Surg.* 198, 105–110 (2004).
24. Mclaughlin, K., Vitale, G., Coderre, S., Violato, C. & Wright, B. Clerkship evaluation–what are we measuring? *Med Teach.* 31, e36-e39 (2009).
25. Gray, J. D. Global rating scales in residency education. *Acad Med.* 71, S55–S63 (1996).
26. Cooper, C. *Individual Differences*. (Routledge, 2002).
27. Winckel, C. P., Reznick, R. K., Cohen, R. & Taylor, B. Reliability and construct validity of a Structured Technical Skills Assessment Form. *Am J Surg.* 167, 423–427 (1994).
28. Scott, D. J. *et al.* Measuring Operative Performance after Laparoscopic Skills Training: Edited Videotape versus Direct Observation. *J. Laparoendosc. Adv Surg Tech.* 10, 183-190 (2000).
29. Sidhu, R. S., Vikis, E., Cheifetz, R. & Phang, T. Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? *Am J Surg.* 191, 677-681 (2006).
30. Melchiors, J. *et al.* Preparing for Emergency: A Valid, Reliable Assessment Tool for Emergency Cricothyroidotomy Skills. *Otolaryngol. -- Head Neck Surg.* 152, 260-265 (2014).
31. Krebs, david e. Opinions and Comments. *Rehabilitation* 67, 22314–22314 (1987).
32. Kottner, J. *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol.* 64, 9096–9106 (2010).
33. Hróbjartsson, A. *et al.* Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *Can Med Assoc J.* 185, E201 (2013).
34. Vogt, V. Y., Givens, V. M., Keathley, C. A., Lipscomb, G. H. & Summitt, R. L. Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? *Am J Obstet Gynecol.* 189, 688-691 (2003).
35. Savović, J. *et al.* Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: Combined analysis of meta-epidemiological studies. *Health Technol Assess. (Rockv).* 16, 1–81 (2012).
36. Kao, L. S., Tyson, J. E., Blakely, M. L. & Lally, K. P. Clinical Research Methodology I: Introduction to Randomized Trials. *J Am Coll Surg.* 206, 361-369 (2008).
37. Korndorffer, J. R., Kasten, S. J. & Downing, S. M. A call for the utilization of consensus standards in the surgical education literature. *Am J Surg.* 199, 99-104 (2010).
38. Vassiliou, M. C. *et al.* Evaluating Intraoperative Laparoscopic Skill: Direct Observation Versus Blinded Videotaped Performances. *Surg Innov.* 14, 211–216 (2007).
39. Schijven, M. P. *et al.* Transatlantic comparison of the competence of surgeons at the start of their professional career. *Br J Surg.* 97, 443–449 (2010).
40. Dath, D. *et al.* Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc.* 18, 1800–1804 (2004).
41. Lievens, F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *J Appl Psychol.* 86, (2001).
42. Matsuda, T. *et al.* Reliability of Laparoscopic Skills Assessment on Video: 8-Year Results of the Endoscopic Surgical Skill Qualification System in Japan. *J Endourol.* 28, 1374–1378 (2014).
43. Gneezy, U. & Rustichini, A. Pay Enough or Don't Pay at All. *Q J Econ.* 115, 791–810 (2000).
44. Cameron, J., Banko, K. M. & Pierce, W. D. Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *Behav Anal.* 24, 1–44 (2001).
45. Frey, B. & Goette, L. Does pay motivate volunteers? 22 (1999). doi:10.3929/ethz-a-004372692
46. Mellström, C. & Johannesson, M. Crowding out in blood donation: Was Titmuss right? *J Eur Econ Assoc.* 6, 845–863 (2008).
47. Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T. & Muijtjens, A. M. M. Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Adv Heal Sci Educ.* 12, 239–260 (2007).