Use of Generalizability Theory for Exploring Reliability of and Sources of Variance in Assessment of Technical Skills: A Systematic Review and Meta-Analysis

Steven Arild Wuyts Andersen, MD, PhD, Leizl Joy Nayahangan, RN, MHCM, Yoon Soo Park, PhD, and Lars Konge, MD, PhD

Abstract

Purpose

Competency-based education relies on the validity and reliability of assessment scores. Generalizability (G) theory is well suited to explore the reliability of assessment tools in medical education but has only been applied to a limited extent. This study aimed to systematically review the literature using G-theory to explore the reliability of structured assessment of medical and surgical technical skills and to assess the relative contributions of different factors to variance.

Method

In June 2020, 11 databases, including PubMed, were searched from inception through May 31, 2020. Eligible studies included the use of G-theory to explore reliability in the context of assessment of medical and surgical technical skills. Descriptive information on study, assessment context, assessment protocol, participants being assessed, and G-analyses was extracted. Data were used to map G-theory and explore variance components analyses. A meta-analysis was conducted to synthesize the extracted data on the sources of variance and reliability.

Results

Forty-four studies were included; of these, 39 had sufficient data for meta-analysis. The total pool included 35,284 unique assessments of 31,496 unique performances of 4,154 participants. Person variance had a pooled effect of 44.2% (95% confidence interval [CI], 36.8%–51.5%). Only assessment tool type (Objective Structured Assessment of Technical Skills-type vs task-based checklist-type) had a significant effect on person variance. The pooled reliability (G-coefficient) was 0.65 (95% CI, .59–.70). Most studies included decision studies (39, 88.6%) and generally seemed to have higher ratios of performances to assessors to achieve a sufficiently reliable assessment.

Conclusions

G-theory is increasingly being used to examine reliability of technical skills assessment in medical education, but more rigor in reporting is warranted. Contextual factors can potentially affect variance components and thereby reliability estimates and should be considered, especially in high-stakes assessment. Reliability analysis should be a best practice when developing assessment of technical skills.

Evidence-based assessment is a cornerstone in competency-based medical education, and reliability is a key aspect of valid assessment scores. In health professions education, assessment is often complex and many variables can contribute to measurement error, including external factors, such as assessors or raters, patient or case variability, procedure difficulty, type of procedure, and interaction with other individuals (e.g., supervisor, surgical team members).[1] Ultimately, assessment comprises "a limited sample of test tasks, measured under unique test conditions,"[2] highlighting the importance of reliable assessment.[3] In medicine and surgery, a large number of assessments are typically not feasible because of a limited caseload for trainees and the lack of experienced assessors that can be physically present during the procedures or assess video recordings of performance. Simulation-based assessment can alleviate some of these challenges but introduces other challenges due to the differences between simulation and real-life conditions, such as the need to transfer skills from one context to another (e.g., from the simulation center to the operating room).[4] Given these challenges, there is a need to integrate multiple types of assessment into the training curriculum to broadly ensure the competency of trainees.[5] More complex ways of measuring competence generally increase reliability by providing a broader mixture of true data (signal) but at the same time can introduce confounders (noise). Consequently, solid statistical methods are necessary to identify sources of noise and to maximize signal, that is, to improve reliability.[2]

Classical test theory can be used to explore reliability (e.g., interrater reliability or test-retest reliability), but it is often not suited to accounting for the complexity of typical assessment conditions in health professions settings. In contrast, generalizability theory (G-theory) allows for robust reliability analysis with the integration of multiple sources and factors contributing to variability of performance and measurement error.[6] In the case of medical and surgical technical skills, this could include variance introduced by the learner's ability, the assessor's leniency, case difficulty, etc. Combining these variance components in the generalizability analysis (G-analysis) allows supplemental decision studies (D-studies) to explore the effects of different combinations of these components on the G-coefficient (e.g., the number of performances

Please see the end of this article for information about the authors.

Correspondence should be addressed to Steven Arild Wuyts Andersen, Copenhagen Academy for Medical Education and Simulation, Blegdamsvej 9, DK-2100, Copenhagen, Denmark; telephone: (+45) 35455054; email: stevenarild@gmail.com; Twitter: @CamesResearch.

Acad Med. 2021;96:1609–1619.
First published online May 4, 2021
doi: 10.1097/ACM.0000000000004150

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/B108.

and assessments in a given assessment context that are necessary to achieve a specific G-coefficient, typically 0.8).[1] G-theory allows not only for the combined reliability under specific assessment conditions to be calculated but also for future assessment approaches that use resources (e.g., trainees' performances, raters, testing modalities) optimally to be devised.

G-theory was developed and refined by Lee Cronbach and colleagues in the 1960s and 1970s.[6] Despite the method's strengths in relation to the analysis of reliability in complex educational settings, it seems that it was the introduction of the objective structured clinical examination (OSCE) format[7] with its use of multiple skills stations and different assessor teams that fueled research into the reliability of assessment in medical education. The Objective Structured Assessment of Technical Skills (OSATS) was introduced much later,[8] prompting the development of a wealth of assessment tools for use in different medical and surgical procedures, including general performance assessment tools, assessment tools for specific procedures, and/or assessment tools for simulation-based assessment.[9] G-theory is equally well suited to explore the reliability of these tools, but it seems that it has only been applied to a limited extent in the literature on structured assessment of technical skills.[9] Consequently, there is limited knowledge on how assessment conditions, such as workplace- and simulation-based assessments, affect generalizability. Furthermore, there is no consensus or even rule of thumb for the ranges of acceptable relative contributions to variance of, for example, learners and assessors in a high-quality structured assessment tool. Finally, G-analyses and reporting seem to vary depending on the authors' and reviewers' experiences and preferences.

The overall aim of this study was to systematically review the literature using G-theory to explore the reliability of structured assessment of medical and surgical technical skills and to assess the relative contributions of different factors (or facets), such as learners, assessors, and assessment context and conditions, to generalizability. Our specific research questions were:

1. How is G-theory being used to assess the reliability of assessments of medical and surgical technical skills?
2. What are the characteristics of the G-analyses and which factors are being explored as contributors to variance?
3. What are the relative contributions of the object of measurement (person variance) and its impact on overall reliability?
4. What is the cutoff for the G-coefficient used in D-studies and how does context affect the optimal performance:observation ratio?

Method

Our systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.[10] The PRISMA statement provides an evidence-based framework for reporting systematic reviews and meta-analyses particularly of randomized, controlled trials and interventional studies. We chose this framework to ensure that we conducted our review with high standards and a systematic approach.

Databases and search strategy

In June 2020, we searched PubMed, Embase, Ovid MEDLINE, Cochrane Library, PLOS Medicine, BioMed Central, OpenGrey, Google Scholar, Directory of Open Access Journals, Scopus, and Web of Science from each database's inception to May 31, 2020. For Google Scholar, we limited the results to the first 70 results for each combination of search strings.

Our search strategy included variations of 3 terms (and their abbreviations or spelling variations)—*generalizability theory*, *generalizability coefficient*, and *decision study*—in combination with any of 4 select key terms—*skills*, *simulation*, *training*, and *surgery*—to narrow the search to the context of technical skills. We piloted these broad search terms in PubMed, and they resulted in all 15 relevant papers from the author group appearing in the searches (the final search string for PubMed is provided in Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/B108). We did not include assessment and technical as search terms as these were not essential in identifying the 15 indicator papers in the pilot and substantially increased the number of nonrelevant papers returned in the results. In contrast, their combination with the other key terms for specificity, such as technical skills or skills assessment, did not result in all of the 15 indicator papers appearing in the results.

Finally, we reviewed the reference lists of identified studies to identify additional potentially eligible studies.

Inclusion and exclusion criteria

Eligible studies included the use of G-theory to explore reliability in the context of assessment of medical and surgical technical skills. To that end, studies that met the following criteria were eligible for inclusion in this systematic review and meta-analysis:

- Population: Studies of medical professionals within medicine or surgery at all learner levels (i.e., medical students, residents, trainees, experienced or experts).
- Intervention: Studies of structured assessment of performance or competency in relation to technical or procedural skills relevant to surgery or medicine.
- Comparison: Studies with educational interventions or observations.
- Outcomes: Studies with reliability outcome measures based on G-analyses.
- Design: Studies with any quantitative design.
- Context: Studies conducted in any health care or medical educational setting reported in any language (but searchable in English).

We excluded:

- Studies that included other health professionals, such as those in dentistry, nursing, and veterinary medicine.
- Studies on nontechnical skills, such as history taking, clinical examination (e.g., auscultation), communication, teamwork, collaboration, general patient management, and feedback skills.
- Studies with combined or aggregated assessment, such as OSCEs and clinical encounters or examinations with a main focus on nontechnical skills or a (subjective) global rating of performance only.

- Abstracts with insufficient quantitative data on G-analyses, theoretical papers, papers discussing only methods, commentaries, and reviews.

Study selection

We saved search results from each database and imported them into the Covidence online platform for systematic reviews (Covidence, Melbourne, Victoria, Australia). Duplicates were removed automatically by the platform or by hand when necessary. Two reviewers (S.A.W.A. and L.J.N.) independently screened titles and abstracts. We included any study deemed potentially relevant by either reviewer for full-text screening. We obtained full texts, and these were screened by both reviewers. Disagreements on final inclusion were resolved by discussion and consensus with the remaining authors (Y.S.P. and L.K.).

Data extraction

We constructed a data extraction form in Excel 2016 (Microsoft Inc., Redmond, Washington), and the 2 reviewers (S.A.W.A. and L.J.N.) piloted this on 5 randomly selected studies. We resolved disagreements by discussion with all authors, and the data from the remaining papers were extracted by 1 reviewer (S.A.W.A.), with the second reviewer (L.J.N.) verifying the extraction of the numerical data from variance analyses. We extracted the following descriptive information from the included studies.

For study information, we extracted the authors, year, country, study design, and study aim.

For assessment context, we extracted the specialty, technical skills or procedure assessed, assessment type (workplace- or simulation-based), assessment modality (clinical, standardized patients, physical models, virtual reality, cadaver, or combinations hereof), name of the assessment tool, assessment tool type (OSATS, task-based checklist [TBC], metrics-based, or combinations hereof), and rating scale (dichotomous, Likert, visual analogue scale score or percentage). We defined OSATS-type assessment tools as being predominately tools structured with individual items rated on a Likert scale, with or without descriptive anchors.[8] We chose to term these tools OSATS-type rather than global assessment tools both to clearly distinguish them from tools consisting of only a single global rating item and because some of these tools are not merely global but adapted to specific procedures or contexts. We defined TBC-type assessment tools as lists of items typically rated performed or not performed, which could be either specific tasks that needed to be performed or errors made during the performance. Finally, metrics-based assessment tools consisted of assessment based on scores derived directly or automatically from simulators or other technical equipment.

For assessment protocol, we extracted the number of procedures per learner, number of assessors per assessment, total number of learners, total number of performances, total number of observations, observation type (live, videorecorded, final product, or combinations hereof), assessor institutional representation (single or multi-institution and if multi-institution, national or international), case variation (yes/no), and assessment tool items (score aggregation, subscoring).

For participants being assessed (i.e., learners), we extracted the learner level (medical students, residents, trainees, experienced or experts). For each of these learner levels, the number of participants, the total number of performances, and the total number of observations were extracted.

For G-analyses, we extracted the G-coefficient cutoff value and reference, method or software used for G-analyses, overall G- or phi-coefficient, and sources of variance (e.g., participants, cases, assessors, items, or any combination hereof reported) in absolute numbers and percentages. If D-studies were performed, we extracted the number of performances and raters (or observations) needed to achieve the chosen cutoff G-coefficient. For papers that did not detail the results of the variance components analysis required for our meta-analysis, we contacted the corresponding author via email and sent 1 reminder after 2 weeks if there was no response.

Data analysis

We used descriptive data for the qualitative synthesis in relation to the research questions and to map the current use of G-theory for analysis of the reliability of assessment of technical skills performance or competency of medical professionals, the chosen cutoff value for the G-coefficient, and which factors are being included in G-analyses. We used the relative contributions of different factors (e.g., participants, assessors, participant*assessor interaction) in the quantitative analysis based on the variance components analyses. In studies where multiple variance components were reported for the same assessment tool (e.g., for different cases or learner levels), we averaged the relative contributions. In cases of multiple variance components being reported but for different assessment tools (e.g., using the same cases or learner levels), we analyzed the data for each assessment tool separately.

We conducted a meta-analysis to synthesize the extracted data on the sources of variance and reliability. Because we aggregated variance components and reliability indices from multiple studies, rather than deriving differences in effect sizes from groups, as is typically done in meta-analysis, we report the adjusted variance components and reliability. When standard error estimates of variance components were unavailable, we used sample-adjusted standard error of measurement to build confidence intervals (CIs).[11] We used a random-effects model to account for heterogeneity in the analyses and checked for model assumptions. In addition, random-effects meta-regression was conducted to examine the impact of assessment setting (workplace- vs simulation-based), assessor institutional representation (single vs multi-institution), observation type (videorecorded vs live), and assessment tool type (OSATS- vs TBC-type). We used the meta set of commands[12] in Stata 16 (StataCorp LLC, College Station, Texas) to pool the results from the identified studies and conduct meta-analyses.

Finally, we assessed the methodological quality of the studies using the Medical Education Research Study Quality Instrument (MERSQI)[13] as operationalized by Cook and Reed.[14] Both reviewers (S.A.W.A. and L.J.N.) appraised the studies in relation to 6 domains: study design, sampling (participant institutional representation and response rate), type of data, validity evidence (content, internal structure, and relationship to other variables), data analysis (sophistication and appropriateness), and outcome. Each

domain is scored from 1 to 3 points using set criteria, resulting in a total score range of 6–18, with higher scores indicating higher quality.

Results

Our search strategy resulted in 3,530 studies after duplicates were removed (Figure 1). Of these, 137 full-text articles were assessed for eligibility with a final inclusion of 44 studies in our qualitative synthesis (see Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/B108).[15–58] For 39 of these studies, sufficient details on the variance components were reported or obtained from the authors for the quantitative synthesis (meta-analysis).[15–27,29–43,45,47–52,54,56–58]

Study characteristics and quality assessment

Overall, the use of G-theory analyses to study aspects of generalizability in technical skills assessment was more frequent in studies published within the last 5 years (2015–2020; Table 1). Further, there seemed to be a high proportion of studies emanating from research environments in Denmark, the United Kingdom, and Canada, totaling >85.0% of the included studies. Unsurprisingly, a majority of the included papers described assessment in surgical specialties (35, 79.5%). We found slightly more papers using G-analyses in the context of workplace-based assessment (28, 63.6%) than simulation-based

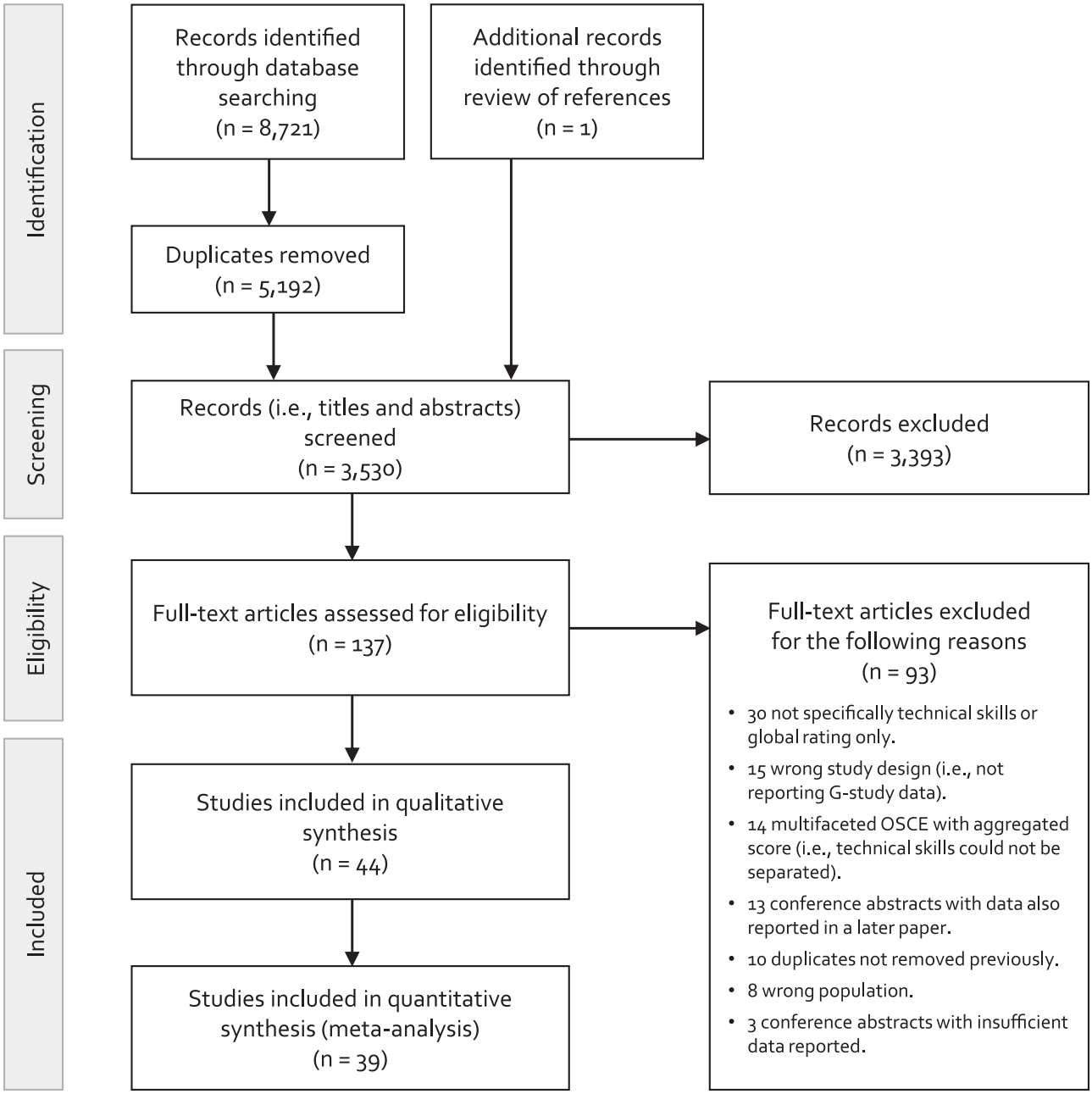

Figure 1 Flowchart showing the selection process for a 2020 systematic review of the literature using generalizability theory to explore the reliability of structured assessment of medical and surgical technical skills and meta-analysis assessing the relative contributions of different factors to variance. Abbreviation: OSCE, objective structured clinical examination.

Table 1

Characteristics of the Included Studies in a 2020 Systematic Review of the Literature Using Generalizability Theory to Explore the Reliability of Structured Assessment of Medical and Surgical Technical Skills and Meta-Analysis Assessing the Relative Contributions of Different Factors to Variance

Study characteristics	No.	%
Total no. of included studies	44	
Year of publication		
Before 2010	2	4.5
2010–2014	17	38.6
2015–2020	25	56.8
Country (primary)		
Denmark	17	38.6
United Kingdom	13	29.5
Canada	8	18.2
The Netherlands	2	4.5
United States	2	4.5
Other	2	4.5
Specialty (primary)		
Surgery	35	79.5
Internal medicine	4	9.1
Anesthesiology	3	6.8
Multiple (both medical and surgical specialties)	2	4.5
Assessment type		
Workplace-based	28	63.6
Simulation-based	16	36.4
Assessment tool type (primary)[a]		
OSATS-type	28	58.3
TBC-type	16	33.3
Metrics-based	2	4.2
Combination of OSATS- and TBC-type	2	4.2
Assessor institutional representation		
Single institution	19	43.2
Multi-institution, national	16	36.4
Multi-institution, international	6	13.6
Not indicated	3	6.8
Simulation-based assessment modality[b]		
Virtual reality	6	33.3
Physical model (e.g., task/box trainer, mannequin)	6	33.3
Standardized patients	2	11.1
Cadaver	1	5.6
Combination of modalities	3	16.7

Abbreviations: OSATS, Objective Structured Assessment of Technical Skills; TBC, task-based checklist.
[a]Some studies included data on several assessment tools, and these are counted separately. Thus, the denominator for this portion of the table is 48.
[b]Two studies on workplace-based assessment used standardized patients as a simulation model. Thus, the denominator for this portion of the table is 18.

assessment (16, 36.4%). Assessors most often represented single institutional assessment (19, 43.2%), followed by multi-institutional assessment with national representation only (16, 36.4%), and then by multi-institutional assessment with international representation (6, 13.6%). In 20 (45.5%) studies, assessment was based on video recordings, in 19 (43.2%) studies—primarily in workplace-based assessment—assessment was based on live observation, and in 2 (4.5%) studies, assessment was based on a combination of the 2 (see Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/B108). The remaining studies based the assessment on metrics (2, 4.5%) and final product analysis (1, 2.3%).

The most frequently reported software used for G-analyses were iterations of the G_String program[59] (16, 36.4%) and the GENOVA/urGENOVA programs[60] (8, 18.2%) or a general statistical software package (18, 40.9%). Two (4.5%) studies did not specify what software was used for statistical analyses.[15,27]

The methodological quality assessment using MERSQI[14] (analysis provided in Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/B108) resulted in a mean score of 14.7 points. Overall, the included studies scored very highly with the sum score for each study ranging from 12.5 to 15.5 out of 18 points.

Characteristics of G-analyses

Altogether, the 44 included studies present data on 4,154 unique participants, 31,496 unique performances, and a total of 35,284 unique assessments of these performances. The median number of performances per learner reported was 2.9 and the median number of assessments/observations per performance reported was 2.0 for the included studies. However, there were large differences in how many unique assessments were contributed by each study. Five studies[31,48–50,57] contributed 27,757 (78.7%) unique assessments; all of these studies were conducted in a workplace-based context with each performance assessed only once. In contrast, there were 11 studies with <30 performances each, which altogether contributed 615 (1.7%) unique assessments.

Most studies (31, 70.5%) reported the variance components for different factors contributing to the reliability of the assessment. For 1 study,[46] variance components analysis was not relevant due to automated, metrics-based assessment (i.e., no assessors) in a workplace setting (i.e., unique cases). The corresponding authors of the remaining 12 studies were contacted, and for 8 studies, variance components analysis data were contributed for this review, resulting in 39 studies included in the quantitative synthesis.

The factors relevant to include in the variance components analysis are highly dependent on study design and the

following overview is provided merely to map which components were most commonly included. Almost all of these 39 studies (37, 94.9%) included variance contributed by participants in the G-analysis with the remaining studies nesting participants within other variables (e.g., case and learner level). This was followed by assessors (32, 82.1%), cases (21, 53.8%), and different combinations of participants, assessors, and cases (i.e., interactions). Finally, some studies (8, 20.5%) included the contribution of individual items in the assessment tooln with or without interactions with participants and assessors. A few studies included some factors that were unique to their study design, for example, rater designation and level of training. All studies included a residual variance component, which is the result of the contributions of any factors that were not included in the analysis. Box plots for the main contributing factors (relative contributions) across all assessment tools and descriptions of the different factors and interactions are presented in Figure 2.

Meta-analysis

In the meta-analysis, which included 39 studies (see above), we focused on the person variance and the overall reliability, as variance components of the G-study design vary by study context and the person variance as the object of measurement was the one that was most consistently included. Person variance and G-coefficient reliability are reported below as pooled effects and as regression effects, as reported in standard meta-analyses approaches.

The participant (person) variance (object of measurement) had a pooled effect of 44.2% (95% CI, 36.8%–51.5%). There was significant evidence of heterogeneity in the effects (I^2 = 95.0%). We chose to explore the effects of the assessment contexts and conditions that allowed sufficient data for the analysis. Testing for difference in pooled effects using meta-regression by setting (workplace- vs simulation-based), assessor institutional representation (single vs multi-institution), observation type (videorecorded vs live), and assessment tool type (OSATS- vs TBC-type) showed that only assessment tool type had a significant effect on the pooled variance component (Table 2). On average, studies with OSATS-type assessment tools had person variance estimates of 48.6% (95% CI, 40.8%–56.3%), in contrast to TBC-type tools, which had person variance estimates of 33.6% (95% CI, 17.8%–49.4%).

The pooled reliability (G-coefficient) was 0.65 (95% CI, .59–.70). There was significant evidence of heterogeneity in the effects (I^2 = 96.6%). Testing for difference in pooled effects using meta-regression showed that there were no significant differences by setting, assessor institutional representation, observation type, or assessment tool type.

Characteristics of D-studies

A majority of the 44 included studies (39, 88.6%) included D-studies to determine the optimal number of performances and raters (or observations) needed to achieve a sufficient G-coefficient. The preferred cutoff for a sufficient G-coefficient was usually 0.8 (24, 54.5%), followed by 0.7 (5, 11.4%), 0.75 (1, 2.3%), and 0.9

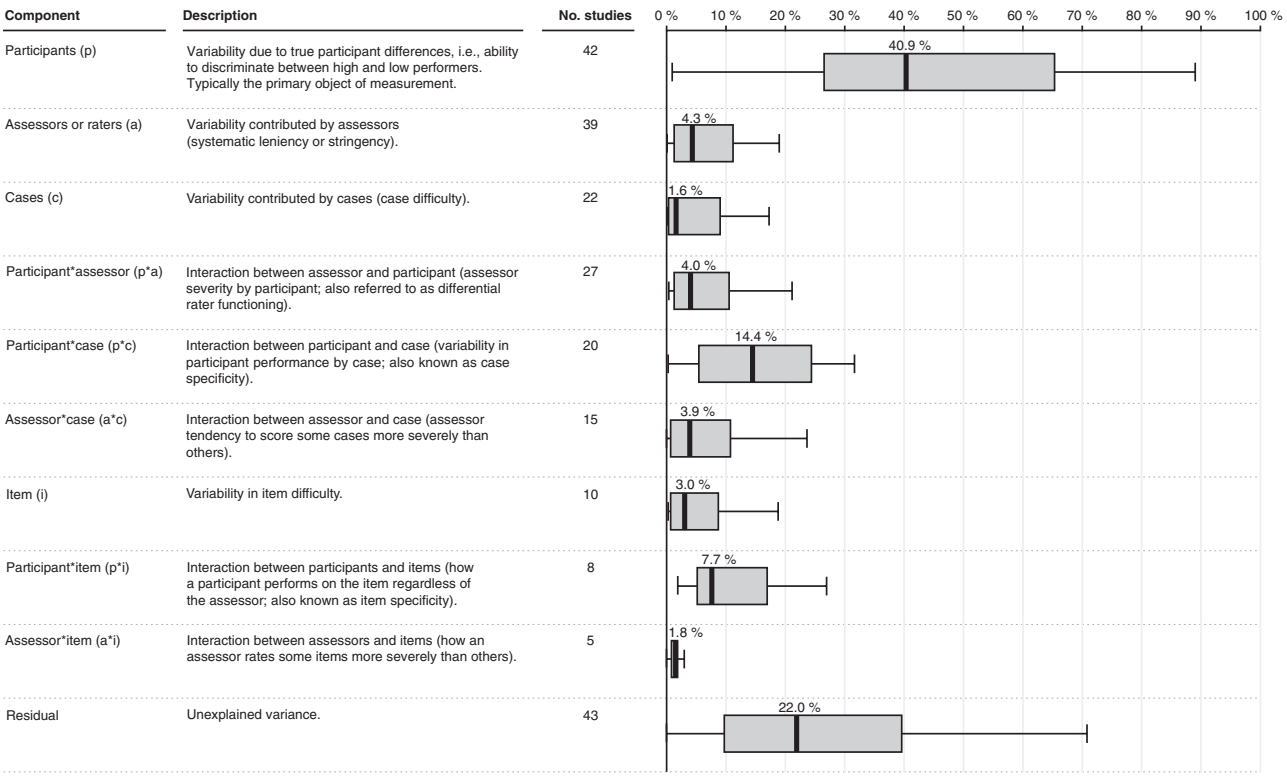

Figure 2 Common factors included in generalizability analyses, with descriptions and box plots, in the 39 studies (with variance components analysis data) included in a 2020 systematic review of the literature using generalizability theory to explore the reliability of structured assessment of medical and surgical technical skills and meta-analysis assessing the relative contributions of different factors to variance. Two of these included studies conducted more than 1 generalizability study, each of which was included in this analysis, thus, the number of studies for some components is greater than 39.

Table 2

Random-Effects Meta-Regression Examining the Impact of Factors on Pooled Effects From the 39 Studies[a] Included in a 2020 Systematic Review of the Literature Using Generalizability Theory to Explore the Reliability of Structured Assessment of Medical and Surgical Technical Skills and Meta-Analysis Assessing the Relative Contributions of Different Factors to Variance

		Participant (person) variance (%)		Reliability (G-coefficient)	
Factor	Comparison[b]	Coefficient (SE)	*P* value	Coefficient (SE)	*P* value
Setting	Workplace- versus simulation-based	–0.05 (0.09)	.596	–0.04 (0.07)	.598
Assessor institutional representation	Single versus multi-institution	0.10 (0.08)	.194	0.10 (0.06)	.100
Observation type	Videorecorded versus live	–0.03 (0.09)	.747	0.05 (0.07)	.465
Assessment tool type	OSATS- versus TBC-type	0.19 (0.10)	.047	0.01 (0.06)	.823
Intercept		0.30 (0.10)	.003	0.60 (0.08)	<.001
Pooled effect		0.44 (0.04)	<.001	0.65 (0.03)	<.001

Abbreviations: SE, standard error; OSATS, Objective Structured Assessment of Technical Skills; TBC, task-based checklist.
[a]With variance components analysis data.
[b]Variables coded as indicator effects. Workplace-based assessment, single-institution assessor, videorecorded observation, and OSATS-type assessment tool coded as 1.

(1, 2.3%). In 13 (29.5%) studies, no G-coefficient cutoff was defined, with 9 of these studies nonetheless performing D-studies. One study defined a cutoff but did not perform D-studies. The cutoff was typically reported to be chosen based on the stakes of the assessment or because the chosen cutoff was reported to be appropriate in a reference paper. The most frequently reported references for the cutoff value were papers by Downing,[3] Crossley et al,[61,62] or Bloch and Norman.[2]

The average ratio of number of performances to assessors to achieve a sufficiently reliable assessment overall was 2.4:1, with similar ratios found for workplace- and simulation-based assessment. For live observations, a higher ratio of performances to assessors seemed to be needed than for videorecorded observations (2.8:1 vs 1.9:1). Similarly, a higher ratio of performances to assessors seemed to be needed for TBC- versus OSATS-type assessment tools (2.9:1 vs 2.1:1) and when assessors represented a single institution rather than multiple institutions (3.0:1 vs 2.1:1).

Discussion

Summary of the main findings

This systematic review and meta-analysis explored the use of G-theory in the context of technical skills assessment in surgery and medicine. Overall, G-theory is increasingly being used to examine the reliability of technical skills assessment in both workplace- and simulation-based settings, but most included studies emanate from just a few countries. Most G-analyses included participants, assessors, cases, and their interactions as factors. The person variance (i.e., the ability to discriminate high and low performers, thereby reflecting true participant differences), which is typically the primary object of measurement, contributed 44.2% (pooled effect). Of the contextual factors, a higher person variance was found for OSATS-type assessment tools compared with TBC-type tools (48.6% vs 33.6%). This indicates that OSATS-type assessment tools might generally be better at capturing true variance. For reliability, a pooled G-coefficient of 0.65 was found, which can be used as a reference for future comparisons of G-coefficients in the technical skills literature. It is notable that the pooled reliability of 0.65 across all 44 studies and 35,284 unique assessments was considerably lower than the most commonly recommended value of 0.8. This implies that both researchers exploring reliability of assessment and clinicians using assessment tools must be prepared to devise more robust assessment setups, including (primarily) more observations of each trainee and (secondarily) more than 1 assessor whenever possible.

Supplementary D-studies were commonly performed to determine the number of performances and assessors needed to achieve reliable assessment with a G-coefficient of ≥0.8. To achieve this, a higher ratio of performances to assessors (i.e., having each learner perform multiple procedures rather than introducing additional assessors) was generally needed. Reliable assessment seemed to require fewer performances and observations in studies where performances were videorecorded or where assessors represented multiple institutions. However, this finding could be confounded and dedicated studies on the effects of different contextual factors on the reliability of assessment are warranted.

Strengths and limitations

Strengths of this systematic review and meta-analysis include the large number of databases searched, even though all publications except for 2 (1 conference abstract[18] and 1 Master's thesis[20]) were indexed in PubMed. To have as much data as possible for the meta-analyses, we contacted all corresponding authors and got additional data for 8 out of the 12 studies that did not report specific details on the variance components.

Some limitations relate to the inclusion and exclusion criteria that were chosen to perform a systematic review in a specific context and to ensure enough comparability of educational context and skill type for a meta-analysis. For example, we limited the procedures to only those performed by medical professionals within medicine and

surgery. Consequently, 2 studies on veterinary students and 1 study on dental students were excluded. Even though we did not a priori exclude basic procedures, such as intravenous catheter insertion or urinary bladder catheterization, the technical skills procedures represented are all taught in postgraduate years (except for maybe lumbar puncture) and only 3 studies included medical students (and used them as a proxy for novice postgraduate trainees).

We also excluded combined or complex assessment situations, such as those performed in OSCEs and clinical encounters for 2 reasons. First, these often include multifaceted assessment of both technical and nontechnical skills, which include communication, diagnostic reasoning, team collaboration, etc. These skills are often difficult to separate and might be more difficult to quantify, potentially affecting interrater reliability. Even though combined technical and nontechnical skills are closer to the ultimate goal of clinical practice, considerably fewer studies combine structured assessment of both. For similar reasons, we also excluded a few studies that had a single measure of technical performance, such as a global rating only, or too few specific items relating to technical performance within a more holistic assessment. Our scope was to explore technical skills assessment and have our results speak specifically to this context. Naturally, this also means that our results cannot be generalized to OSCE-type assessments and assessments combining both technical and nontechnical elements. Second, study designs are often complex in OSCEs, which can result in unbalanced or hierarchical study designs and observations being nested within assessor groups and stations, examination day, or site. This adds more facets to the G-analyses, making direct comparison with simpler assessment situations increasingly difficult. Despite aiming for comparability for our meta-analysis, the different study designs of the included studies made it challenging to synthesize the pooled effect for different variance components. However, we were able to obtain the pooled effects for person variance and overall reliability, which have the most impact on the overall quality of assessment scores.

Comparison with other studies

This is the first systematic review and meta-analysis on the use of G-theory within medical education. A 2011 systematic review on observational tools for assessment of procedural skills included only a single study that used G-theory for exploring reliability.[9] In contrast, we identified 44 studies on technical skills assessment with G-analyses and used these data to map the characteristics of how G-theory is used, to map the relative contributions of different sources of variance to reliability, and to estimate contextual factors that affect generalizability of technical skills assessment, about which little is known. Only a single previous study from 2005 investigated how context influenced reliability for a general resident performance rating instrument and reported that reliability increased as assessors (and potentially also the learners) gained more experience with the assessment tool (i.e., reliability increased after the first year of using the tool).[63]

There is some knowledge on the effects on reliability of assessment tool type and observation context. We found indications that TBC-type assessment tools require a higher performance to assessor ratio than OSATS-type assessment tools. This is in agreement with the findings by Regehr et al, who compared the psychometric properties of TBC- versus OSATS-type assessment tools and found that TBC-type tools had an inferior ability to discriminate learner level and to predict operative outcomes compared with OSATS-type tools.[64] It is also in agreement with findings from a systematic review by Ilgen et al that found global-rating type instruments, such as the OSATS, have a higher average interitem and interstation reliability in simulation-based assessment than checklists.[65] For other factors relating to the context of technical skills assessment, there are few studies. For type of observation, a higher interrater reliability (intraclass correlation coefficient) of direct observation (0.99, 2 raters) compared with video raters (0.68, 4 raters) for the Global Operative Assessment of Laparoscopic Skills (GOALS) assessment tool has been reported.[66] In contrast, we found that more performances were generally needed for reliable assessment of live performances than for video recordings. This highlights the need for dedicated studies on the effects of context on the reliability of assessment.

The included studies were all rated highly on the MERSQI tool. This was in large part related to the inclusion criteria of this systematic review, such as quantitative study design and the use of G-theory in the outcome. Consequently, all studies reported objective assessment data (i.e., type of data = 3 points on the MERSQI tool), data analysis (i.e., sophistication beyond descriptive analysis = 2 points and data analysis appropriate = 1 point), and outcome (i.e., studies reporting on skills as an outcome = 1.5 points), and all studies provided validity evidence for the assessment (i.e., internal structure = 1 point). Furthermore, most studies included considerations on content validity of the assessment tool (1 point) and relationship to other variables, such as training level or experience (1 point); had data sampling with participants representing multiple institutions (1.5 points) and high response rates (1.5 points); and included 2 or more groups (2 points). It is therefore unsurprising that the included papers in our review scored substantially higher (mean = 14.7) than what Cook and Reed found across all the reviews included in their study (mean = 11.3 points).[14]

Implications

Satisfactory reliability of assessment, especially in higher-stakes assessment, is very important and G-theory offers advantages over classical test theory and the use of, for example, Cronbach's alpha, as a measure of reliability. Cronbach and Shavelson later wrote that the alpha coefficient "[…] is now seen to cover only a small perspective of the range of measurement uses for which reliability information is needed."[67] In contrast, G-theory can be used to explore and optimize factors, such as assessors, cases, and assessment tool items, contributing to reliability for the specific assessment situation. We found person variance, which is typically desired to be the highest contributor to the score variance, accounted for less than 50% of variance. This varied greatly between the individual included studies (from 3.7%[47] to 89.0%[17]), but the pooled person variance (44.2%) can provide a point of reference for judging new assessment tools. Further, we

found that OSATS-type tools are better at capturing person variance than TBC-type assessment tools, supplementing other studies that have also found OSATS-type assessment tools to have better psychometric properties than checklists.[64,65] When developing new tools for the assessment of surgical and medical technical skills, choosing an OSATS-type tool will likely result in higher reliability, but systematic evaluation should be mandated for any new tool.

In relation to research, it is important that G-theory is applied with high methodological rigor. In a review of methodological trends in G-theory, Rios et al found 58 studies applying G-theory in an educational and/or psychological setting.[68] They found that more than half of the included studies had a small sample size (<100 observations) and that many studies imputed missing data to use a balanced design for the analysis. Further, they found that only a few studies reported standard error estimates of the variance components as many statistical packages do not provide these in their standard packages, which is concerning, especially in light of the studies often having small sample sizes.[68] Even though we found the included papers in our systematic review to be high quality using the MERSQI tool, there was limited information on the G-analysis and estimates, suggesting that methodological rigor in analyses is also a concern in medical education. Crossley et al have made recommendations for designing, conducting, and reporting the results of G-analyses in medical educational research.[69] These include ensuring adequate sampling of all factors relevant for reliability, detailing the statistical procedures used, and presenting raw values for the variance components, percentage of variance, and degrees of freedom for each facet. Based on our experience with the data collection for this systematic review, we would further recommend that papers clearly state the number of participants (preferably according to learner levels), assessment protocol (i.e., planned number of performances and observations or assessors for each performance as well as the actual totals for each of these), which factors were considered, how factors were nested, whether the study design was balanced or unbalanced, what software was used for the G-analyses, and if a D-study was performed, which G-coefficient cutoff was used and the resulting optimal performance:assessor ratio to achieve this.

Studies using G-theory to investigate reliability of structured assessment of medical and surgical technical skills are mostly from within the last 10 years and frequently emanate from a few research environments in Denmark, the United Kingdom, and Canada. The reason for this is most likely that there are active research groups examining simulation- or workplace-based assessment in these areas, and in the case of the United Kingdom, there is national implementation of systematic competency assessment in endoscopy with reporting to central databases.[48] To ensure that G-theory is more widely used in medical education research, it is important that current experts share their knowledge of and experience with G-theory methods with other groups. There are several excellent papers in the medical education literature that detail how to use G-theory with hands-on examples of how to conduct G-analyses and relevant considerations.[2,61] On an encouraging note, we found that most of the conference abstracts that included G-theory methods found in our searches were later published as full papers (13/14, 93%). This should encourage the use of G-theory methods in establishing validity evidence of assessments because the value added potentially contributes to a high publication rate.

Ultimately, this systematic review and meta-analysis suggests that more research is needed to investigate how contextual factors, such as observation type and assessor institutional representation, contribute to reliability and to understand the reasons for this.

Conclusions

In this systematic review and meta-analysis, we found that G-theory is increasingly being used to study the reliability of structured assessment in surgical and medical technical skills training and that some contextual factors potentially affect person variance. This could potentially affect reliability estimates, and even though we did not find a statistically significant effect on reliability of assessment in our meta-analysis, this should still be considered, especially in high-stakes assessment. Altogether, reliability analysis should be a best practice when developing assessment of technical skills. Ultimately, there is a need for more dedicated studies investigating the effects of context on reliability.

Funding/Support: S.A.W. Andersen has received research funding for his postdoctoral research from the Independent Research Fund Denmark (8026-00003B). The remaining authors have no other sources of funding or support to declare.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

Data: Data and the study protocol can be requested from the corresponding author.

S.A.W. Andersen is postdoctoral researcher, Copenhagen Academy for Medical Education and Simulation (CAMES), Center for Human Resources and Education, Capital Region of Denmark, and Department of Otolaryngology, The Ohio State University, Columbus, Ohio, and resident in otorhinolaryngology, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Copenhagen, Denmark; ORCID: https://orcid.org/0000-0002-3491-9790.

L.J. Nayahangan is researcher, CAMES, Center for Human Resources and Education, Capital Region of Denmark, Copenhagen, Denmark; ORCID: https://orcid.org/0000-0002-6179-1622.

Y.S. Park is director of health professions education research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts; ORCID: https://orcid.org/0000-0001-8583-4335.

L. Konge is professor of medical education, University of Copenhagen, and head of research, CAMES, Center for Human Resources and Education, Capital Region of Denmark, Copenhagen, Denmark; ORCID: https://orcid.org/0000-0002-1258-5822.

References

1 Bilgic E, Watanabe Y, McKendy KM, Ito Y, Vassiliou MC. Reliable assessment of performance in surgery: A practical approach to generalizability theory. J Surg Educ. 2015;72:774–775.
2 Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE guide no. 68. Med Teach. 2012;34:960–992.
3 Downing SM. Reliability: On the reproducibility of assessment data. Med Educ. 2004;38:1006–1012.
4 Haskell RE. Transfer of learning: What it is and why it's important. In: Transfer of Learning: Cognition, Instruction, and Reasoning. San Diego, CA: Academic Press; 2001:23–39.
5 Schuwirth LW, Southgate L, Page GG, et al. When enough is enough: A conceptual basis for fair and defensible practice performance assessment. Med Educ. 2002;36:925–930.
6 Brennan RL. Generalizability theory and classical test theory. Appl Measurement Educ. 2011;24:1–21.
7 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13:41–54.

8 Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273–278.
9 Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: A systematic review. Am J Surg. 2011;202:469–480.e6.
10 Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. PLOS Med. 2009;6:e1000097.
11 Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods. 2010;1:39–65.
12 Palmer TM, Sterne JAC. Meta-Analysis in Stata: An Updated Collection from the Stata Journal, 2nd ed. College Station, TX: Stata Press; 2016.
13 Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA. 2007;298:1002–1009.
14 Cook DA, Reed DA. Appraising the quality of medical education research methods: The Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med. 2015;90:1067–1076.
15 Barton JR, Corbett S, van der Vleuten CP; English Bowel Cancer Screening Programme; UK Joint Advisory Group for Gastrointestinal Endoscopy. The validity and reliability of a Direct Observation of Procedural Skills assessment tool: Assessing colonoscopic skills of senior endoscopists. Gastrointest Endosc. 2012;75:591–597.
16 Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: A prospective observational study of the methodology. Health Technol Assess. 2011;15:1–194.
17 Bech B, Lönn L, Falkenberg M, et al. Construct validity and reliability of structured assessment of endoVascular expertise in a simulated setting. Eur J Vasc Endovasc Surg. 2011;42:539–548.
18 Bilgic E, Watanabe Y, Lee L, Vassiliou MC. Reliability of GOALS scores using generalizability theory. Conference abstract and poster presented at: Society of American Gastrointestinal and Endoscopic Surgeons (SAGES); April 2–5, 2014; Salt Lake City, UT. https://www.sages.org/meetings/annual-meeting/abstracts-archive/reliability-of-goals-scores-using-generalizability-theory. Accessed April 9, 2021.
19 Bilgic E, Watanabe Y, McKendy K, et al. Reliable assessment of operative performance. Am J Surg. 2016;211:426–430.
20 Budden CR. Using the Ottawa Surgical Competency Operative Room Evaluation (O-SCORE) in a Canadian Plastic Surgery Program [Master's thesis]. Edmonton, Alberta, CA: University of Alberta; 2016.
21 Carlsen CG, Lindorff-Larsen K, Funch-Jensen P, Lund L, Charles P, Konge L. Reliable and valid assessment of Lichtenstein hernia repair skills. Hernia. 2014;18:543–548.
22 de Vries AH, Muijtjens AMM, van Genugten HGJ, et al. Development and validation of the TOCO-TURBT tool: A summative assessment tool that measures surgical competency in transurethral resection of bladder tumour. Surg Endosc. 2018;32:4923–4931.
23 Fernandez SA, Wiet GJ, Butler NN, Welling B, Jarjoura D. Reliability of surgical skills scores in otolaryngology residents: Analysis using generalizability theory. Eval Health Prof. 2008;31:419–436.
24 Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
25 Graeser K, Konge L, Kristensen MS, Ulrich AG, Hornbech K, Ringsted C. Airway management in a bronchoscopic simulator based setting: An observational study. Eur J Anaesthesiol. 2014;31:125–130.
26 Guldbrand Nielsen D, Jensen SL, O'Neill L. Clinical assessment of transthoracic echocardiography skills: A generalizability study. BMC Med Educ. 2015;15:9.
27 Gupta S, Anderson J, Bhandari P, et al. Development and validation of a novel method for assessing competency in polypectomy: Direct observation of polypectomy skills. Gastrointest Endosc. 2011;73:1232–1239.e2.
28 Harris A, Butterworth J, Boshier PR, et al. Development of a reliable surgical quality assurance system for 2-stage esophagectomy in randomized controlled trials [published online ahead of print March 27, 2020]. Ann Surg. doi:10.1097/SLA.0000000000003850.
29 Henriksen MJV, Wienecke T, Thagesen H, et al. Assessment of residents readiness to perform lumbar puncture: A validation study. J Gen Intern Med. 2017;32:610–618.
30 Hertz P, Jensen K, Abudaff SN, et al. Ensuring basic competency in chest tube insertion using a simulated scenario: An international validation study. BMJ Open Respir Res. 2018;5:e000362.
31 Homer M, Setna Z, Jha V, Higham J, Roberts T, Boursicot K. Estimating and comparing the reliability of a suite of workplace-based assessments: An obstetrics and gynaecology setting. Med Teach. 2013;35:684–691.
32 Jensen K, Hansen HJ, Petersen RH, et al. Evaluating competency in video-assisted thoracoscopic surgery (VATS) lobectomy performance using a novel assessment tool and virtual reality simulation. Surg Endosc. 2019;33:1465–1473.
33 Kara CO, Mengi E, Tümkaya F, Ardıç FN, Şenol H. Adaptation of "Objective Structured Assessment of Technical Skills" for adenotonsillectomy into Turkish: A validity and reliability study. Turk Arch Otorhinolaryngol. 2019;57:7–13.
34 Konge L, Arendrup H, von Buchwald C, Ringsted C. Using performance in multiple simulated scenarios to assess bronchoscopy skills. Respiration. 2011;81:483–490.
35 Konge L, Larsen KR, Clementsen P, Arendrup H, von Buchwald C, Ringsted C. Reliable and valid assessment of clinical bronchoscopy performance. Respiration. 2012;83:53–60.
36 Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C. Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. Endoscopy. 2012;44:928–933.
37 Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C. Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration. 2013;86:59–65.
38 Konge L, Clementsen PF, Ringsted C, Minddal V, Larsen KR, Annema JT. Simulator training for endobronchial ultrasound: A randomised controlled trial. Eur Respir J. 2015;46:1140–1149.
39 Lord JA, Zuege DJ, Mackay MP, des Ordons AR, Lockyer J. Picking the right tool for the job: A reliability study of 4 assessment tools for central venous catheter insertion. J Grad Med Educ. 2019;11:422–429.
40 MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the simulated environment. Teach Learn Med. 2016;28:72–79.
41 Marriott J, Purdie H, Crossley J, Beard JD. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. Br J Surg. 2011;98:450–457.
42 McLeod G, McKendrick M, Taylor A, et al. Validity and reliability of metrics for translation of regional anaesthesia performance from cadavers to patients. Br J Anaesth. 2019;123:368–377.
43 Melchiors J, Petersen K, Todsen T, Bohr A, Konge L, von Buchwald C. Procedure-specific assessment tool for flexible pharyngo-laryngoscopy: Gathering validity evidence and setting pass-fail standards. Eur Arch Otorhinolaryngol. 2018;275:1649–1655.
44 Miskovic D, Ni M, Wyles SM, et al; National Training Programme in Laparoscopic Colorectal Surgery in England. Is competency assessment at the specialist level achievable? A study for the national training programme in laparoscopic colorectal surgery in England. Ann Surg. 2013;257:476–482.
45 Moiz B, Ali SK, Rashid A, Shariq M, Karim F. Development and pilot testing of a novel tool for evaluating practical skills in hematopathology residents in Pakistan. J Grad Med Educ. 2019;11(suppl 4):177–180.
46 Preisler L, Søndergaard Svendsen MB, Søndergaard B, et al. Automatic and unbiased assessment of competence in colonoscopy: Exploring validity of the Colonoscopy Progression Score (CoPS). Endosc Int Open. 2016;4:E1238–E1243.
47 Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: Assessing technical and non-technical skills of internal medicine residents. Adv Health Sci Educ Theory Pract. 2015;20:85–100.
48 Siau K, Crossley J, Dunckley P, et al. Training and assessment in flexible sigmoidoscopy: Using a novel direct observation of procedural skills (DOPS) assessment tool. J Gastrointestin Liver Dis. 2019;28:33–40.
49 Siau K, Crossley J, Dunckley P, et al; Joint Advisory Group on Gastrointestinal Endoscopy (JAG). Direct observation of procedural skills (DOPS) assessment in diagnostic gastroscopy: Nationwide evidence of validity and competency development during training. Surg Endosc. 2020;34:105–114.

50 Siau K, Crossley J, Dunckley P, et al; Joint Advisory Group on Gastrointestinal Endoscopy (JAG). Colonoscopy direct observation of procedural skills assessment tool for evaluating competency development during training. Am J Gastroenterol. 2020;115:234–243.

51 Strøm M, Lönn L, Konge L, et al. Assessment of EVAR competence: Validity of a Novel Rating Scale (EVARATE) in a simulated setting. Eur J Vasc Endovasc Surg. 2018;56:137–144.

52 Thomsen AS, Bach-Holm D, Kjærbo H, et al. Operating room performance improves after proficiency-based virtual reality cataract surgery training. Ophthalmology. 2017;124:524–531.

53 Tjiam IM, Schout BMA, Hendrikx AJM, et al. Program for laparoscopic urological skills assessment: Setting certification standards for residents. Minim Invasive Ther Allied Technol. 2013;22:26–32.

54 Todsen T, Tolsgaard MG, Olsen BH, et al. Reliable and valid assessment of point-of-care ultrasonography. Ann Surg. 2015;261:309–315.

55 Tsai AY, Mavroveli S, Miskovic D, et al. Surgical quality assurance in COLOR III: Standardization and competency assessment in a randomized controlled trial. Ann Surg. 2019;270:768–774.

56 Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42:364–373.

57 Williams RG, Verhulst S, Colliver JA, Sanfey H, Chen X, Dunnington GL. A template for reliable assessment of resident operative performance: Assessment intervals, numbers of cases and raters. Surgery. 2012;152:517–524.

58 Winkler-Schwartz A, Marwa I, Bajunaid K, et al. A comparison of visual rating scales and simulated virtual reality metrics in neurosurgical training: A generalizability theory study. World Neurosurg. 2019;127:e230–e235.

59 Bloch R, Norman G. Generalizability theory tool: G_String program. McMaster University. https://healthsci.mcmaster.ca/merit/research/g_string_v. Accessed April 12, 2021.

60 Crick JE, Brennan RL. Computer programs: GENOVA/urGENOVA programs. University of Iowa College of Education. https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs. Accessed April 12, 2021.

61 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. Med Educ. 2002;36:972–978.

62 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45:560–569.

63 Williams RG, Verhulst S, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: More items or more observations? Surgery. 2005;137:141–147.

64 Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73:993–997.

65 Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49:161–173.

66 Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: Direct observation versus blinded videotaped performances. Surg Innov. 2007;14:211–216.

67 Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Measurement. 2004;64:391–418.

68 Rios JA, Li X, Faulkner-Bond M. A review of methodological trends in generalizability theory. Paper presented at: Northeastern Educational Research Association Conference; October 2013; Rocky Hill, CT. https://www.researchgate.net/publication/281034264_A_Review_of_Methodological_Trends_in_Generalizability_Theory. Accessed April 12, 2021.

69 Crossley J, Russell J, Jolly B, et al. 'I'm pickin' up good regressions': The governance of generalisability analyses. Med Educ. 2007;41:926–934.