ACADEMIC MEDICINE
Journal of the Association of American Medical Colleges

Author: Andersen Steven Arild Wuyts MD, PhD; Park Yoon Soo PhD; Sørensen Mads Sølvsten MD, DMSc; Konge Lars MD, PhD

Title: Reliable Assessment of Surgical Technical Skills Is Dependent on Context: An Exploration of Different Variables Using Generalizability Theory

DOI: 10.1097/ACM.0000000000003550

Academic Medicine

DOI: 10.1097/ACM.0000000000003550

Reliable Assessment of Surgical Technical Skills Is Dependent on Context: An Exploration of Different Variables Using Generalizability Theory

Steven Arild Wuyts Andersen, MD, PhD, Yoon Soo Park, PhD, Mads Sølvsten Sørensen, MD, DMSc, and Lars Konge, MD, PhD

S.A.W. Andersen is postdoc, Copenhagen Academy for Medical Education and Simulation (CAMES), Center for HR & Education, the Capital Region of Denmark, and otorhinolaryngology resident, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Copenhagen, Denmark; ORCID: http://orcid.org/0000-0002-3491-9790.

Y.S. Park is associate professor, Department of Medical Education, University of Illinois – College of Medicine at Chicago, Chicago, Illinois; ORCID: http://orcid.org/0000-0001-8583-4335.

M. S. Sørensen is professor of otorhinolaryngology, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Copenhagen, Denmark, and head of the Visible Ear Simulator project.

L. Konge is professor of medical education, University of Copenhagen, Denmark, and head of research, Copenhagen Academy for Medical Education and Simulation (CAMES), Center for HR & Education, the Capital Region of Denmark.

Correspondence should be directed to: Dr. Steven Andersen, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark; telephone: +45 35452072; email: stevenarild@gmail.com; Twitter (institution): @RegionH.

Acknowledgements: The authors would like to acknowledge Professor Per Cayé-Thomasen and Dr. Søren Foghsgaard for their substantial contributions to rating the performances throughout the years, and Mr. Peter Trier Mikkelsen who programmed the Visible Ear Simulator.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: The regional ethics committee for the Capital Region of Denmark had deemed exempt each of the individual studies from which the authors pulled data.

Previous presentations: This work was presented as a short communication at the 2019 Association for Medical Education in Europe (AMEE) Conference, August 25-28, 2019, Vienna, Austria.

Abstract

Purpose

Reliable assessment of surgical skills is vital for competency-based medical training. Several factors influence not only the reliability of judgements but also the number of observations needed for making judgments of competency that are both consistent and reproducible. The aim of this study was to explore the role of various conditions—through the analysis of data from large-scale, simulation-based assessments of surgical technical skills—by examining the effects of those conditions on reliability using Generalizability theory.

Method

Assessment data from large-scale, simulation-based temporal bone surgical training research studies in 2012–2018 were pooled, yielding collectively 3,574 assessments of 1,723 performances. The authors conducted generalizability analyses using an unbalanced random-effects design, and they performed decision studies to explore the effect of the different variables on projections of reliability.

Results

Overall, five observations were needed to achieve a Generalizability coefficient > 0.8. Several variables modified the projections of reliability: increased learner experience necessitated more observations (5 for medical students, 7 for residents, and 8 for experienced surgeons); the more complex cadaveric dissection required fewer observations than virtual reality simulation (2 vs. 5 observations); and increased fidelity simulation graphics reduced the number of observations needed from 7 to 4. The training structure (either massed or distributed practice) and simulator-integrated tutoring had little effect on reliability. Finally, more observations were needed during initial training when the learning curve was steepest (6 observations) compared with the plateau phase (4 observations).

Conclusions

Reliability in surgical skills assessment seems less stable than it is often reported to be. Training context and conditions influence reliability. The findings from this study highlight that medical educators should exercise caution when using a specific simulation-based assessment in other contexts.

Valid and reliable assessment of performance is vital for competency-based surgical training.[1] Any assessment represents "a limited sample of test tasks, measured under unique test conditions to a universe of tasks and conditions, from which the specific test set has been drawn more or less arbitrarily."[2] Ideally, therefore, the assessment of surgical technical skills should be based on a large number of observations to ensure reliability.[3] A sufficient number of observations to achieve adequate reliability, however, is often not feasible if procedural time is lengthy. First, the number of procedures a single learner can perform is limited, and the costs of direct observation or blinded assessment of videotaped procedures by several external assessors is high.[4] Furthermore, in surgical skills assessment, as in the assessment of skills within all medical specialties, many variables contribute to measurement error including external factors such as raters, patient variability, the complexity of the procedure, the type of procedure, and interactions with other individuals.[5]

In contrast to traditional approaches to estimating reliability such as classical test theory, Generalizability theory (G theory) integrates multiple sources of factors contributing to the variability of performance and measurement error, thereby allowing a more robust reliability analysis of complex assessment methods.[6] A generalizability analysis can be performed to extract the G-coefficient and in surgical technical skills assessment, a G-coefficient of > 0.8 is often considered acceptable.[7] The generalizability analysis can also be used to explore and determine the optimal number of raters and performances for reliable assessment with a number of supplemental decision studies (D studies).[8] This results in the reliability often being expressed as the number of observations (number of raters and performances) for assessment needed to achieve a G-coefficient > 0.8.

These analyses are typically based on data from single studies that represent a specific assessment context, which is often not considered when using the specific assessment tool in other contexts—for example, at other institutions or with other modalities (e.g., using a tool developed for assessment during simulation-based training for, instead, assessment during

real-life surgery). Importantly, context could influence assessment reliability due to varying score distributions, and a reported G-coefficient cannot be assumed to be a general trait of the assessment tool.

Medical educators from all specialties, including otorhinolaryngology (ORL), recognize the need for high-quality assessment of competency, which is especially important in potentially high-stakes assessment (e.g., board certification).[9,10] Simulation-based assessment of technical skills reduces some of the real-life clinical variability and allows assessment to occur in a controlled environment. Simulation-based assessment could, therefore, potentially minimize some of the sources of error in technical skills assessment, which in turn, would allow investigators to explore the remaining factors contributing to measurement error. Thus far, however, few investigations have explored the effects of different variables on the reliability of simulation-based assessment of surgical skills. In ORL in general, and in temporal bone surgery specifically,[11,12] there remains a gap in the systematic implementation of simulation-based training into the curricula despite favorable evidence of its effectiveness.[13] One potential barrier could be that, heretofore, the field has lacked well-defined levels of performance (e.g., standard setting for use in mastery learning).[14,15]

The availability of a large amount of assessment data from temporal bone surgical training afforded us the opportunity to apply G theory to the data to investigate several assessment context variables. Our goal was to answer the following research question: What are the effects of different contexts and conditions on reliability of simulation-based assessment?

Method

Data

We pooled all assessment data from simulation-based temporal bone surgical training research studies at our institution that occurred from 2012 to 2018[15–24] (Figure 1). The procedure performed in all the training session—the principal temporal bone procedure—was the same: complete anatomical mastoidectomy ± posterior tympanotomy (the further drilling

of the small bony plate between the facial nerve and chorda tympani). Raters used the modified Welling Scale,[16] a final-product analysis tool, for all assessments. The tool consists of 25 or 26 items (depending on ± posterior tympanotomy) rated dichotomously (either 0 or 1). These items reflect key steps/objectives of the temporal bone drilling (e.g., the identification of the vertical part of the facial nerve without causing injury to the nerve). Four raters experienced with teaching temporal bone surgery performed all assessments across all the studies; they were blinded to learner identity and level, procedure number, training structure, and the use of tutoring.

We examined 2 modalities of simulation-based training: training on human cadavers (dissection) and training on a virtual reality (VR) simulator. The temporal bone of the human cadavers provide natural anatomical variation. The laboratory facilities of the Department of Anatomy, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark, provided the dissection set up, which comprised the heads of fresh frozen cadavers (donated material), standard operating microscopes, and an otosurgical drill with a range of drill bits and suction/irrigation functionality. Only residents performed dissection on cadavers, and each learner performed only one dissection procedure.

The Visible Ear Simulator (VES) is an established temporal bone surgical simulator for VR simulation training of mastoidectomy.[25] It is available as academic freeware[26] and runs on PC [personal computer] platforms with a Geforce GTX graphics card (Nvidia Corp, Santa Clara, California) and a Geomagic Touch haptic device (3D Systems Inc., Rock Hill, South Carolina). The device allows for interaction and provides force feedback during drilling. The VR simulator also features an integrated tutor function that greenlights the volume (i.e., amount) to be drilled in each step of the procedure; the system thereby provides a visual cue indicating where to drill but also allows learners to drill outside the defined volume without alerting them.

The VR simulation used only a single virtual temporal bone model, but the bone was rendered in three different qualities depending on the version of the simulator (versions 1.3, 2.1, and 3.0, each representing successively increasing graphic fidelity). Unlike in the cadaveric dissection, in VR simulation, learners from three different levels participated: medical students, ORL residents, and experienced otosurgeons. Learners completed between 1 and 18 VR simulation procedures. Furthermore, both the training structure (distributed or massed practice) and learning support varied (with or without greenlighting [i.e., simulator integrated tutoring]) across the VR simulation procedures. We investigated the effects of this greenlighting tutoring at both the level of the procedure itself (i.e., tutoring during the procedure) and at the level of the entire training program (i.e., tutoring during training). That is, we examined the effect of receiving greenlighting tutoring during the assessment vs. no tutoring during the assessment, and we compared learners who had received greenlighting tutoring at any point in their training vs. those who had never received any tutoring.

Finally, in the VR simulation, we studied different parts of the learning curve of repeated practice based on the average learning curve[20]: the initial phase (procedures 1–5) that has a steep slope (most learners will demonstrate a rapid increase in performance); an interim phase (procedures 6–10) that has a shallower average slope (representing a mix of some learners who continue to improve their performance whereas others have reached their potential); and a plateau phase (procedure 11–15) that is nearly flat (most learners have reached their potential).

Statistics

We stratified data according to training variables (subgroups), including learner level (medical student, ORL resident, experienced surgeon) and simulation modality (dissection vs. VR simulation). Other subgroup variables for the VR simulated procedures included the fidelity of the graphics (simulator version 1.3, 2.1, or 3.0), the training structure (massed or distributed practice), VR simulator-integrated tutoring during the procedure (vs. without

greenlighting during the procedure), VR simulator-integrated tutoring during training (vs. no tutoring at any time during training), and the learning curve phase (initial, interim, or plateau). We considered all of these subgroups to be factors for stratification in our variance components estimations. We used descriptive statistics to examine trends between different variables and raters.

We conducted generalizability analyses for the full sample and also for each subgroup outlined above using urGENOVA.[27] Given the unbalanced nature of the data (different number of observations per learner), we used the unbalanced random-effects design, as suggested by Brennan and applied in the medical educational literature[28,29]:

[observation (o) : person (p)] x item (i).

We sorted data by the timing of the assessment (occasion), to capture changes in variability. We included all performances from a particular data set for each subgroup; for example, we analyzed all performances with different simulator versions in our G study of graphics fidelity. We used variance components from the G study to examine sources of error variance and the reliability of the assessment. Following each of the G studies, we performed D studies to explore the effects on projected reliability (the G-coefficient provides reliability estimates for use when making normative decisions). We compared D study results between subgroups, examining projections in reliability. We set a G-coefficient of 0.8 as the threshold for reliable assessment, and we reported the number of observations needed to reach this threshold in a context of the specified variables. We used Stata 14 (StataCorp LLC, College Station, Texas) for data compilation and analyses.

Ethics

The regional ethics committee for the Capital Region of Denmark had deemed exempt each of the individual studies from which we pulled data. All participants volunteered for the studies, all signed informed consent after receiving thorough information about the studies, and none received compensation for participation.

Results

We included a total of 3,574 assessments of 1,723 unique performances by 246 participants across 3 learner levels: 103 medical students, 132 ORL residents, and 11 experienced otosurgical experts. Four experienced raters contributed to the pool of assessments. See also Table 1.

Generalizability study

We used an unbalanced random-effects generalizability study to decompose variance components of the data, following this design: observation nested in person crossed with items. Person variance (object of measurement; true variance) was 4.8%, indicating that the assessment tool was able to discriminate between high and low performers. The largest source of variability was due to items (36.9%), implying variability in item difficulty (some items were more difficult than others). The rater effect, as measured using observation nested in person, was 4.6%, signaling consistency in raters' assessment of the encounter (low variability in rater stringency). We detected modest evidence of item specificity (10.8%), indicating that some learners who performed well on certain items did not perform well on other items (person-item interaction). Using these data configurations, the G-coefficient reliability was .80 with standard error of measurement of .05. We have presented the variance components (estimate and percent), degrees of freedom, and an interpretation in Table 2.

Decision study

Results from the D study on the overall dataset showed that the number of observations needed to achieve a G-coefficient > 0.8 was 5 (Figure 2); however, many of the different training variables changed the general number of observations needed (Figure 3A–G). To examine projections of reliability using G theory, we replicated the unbalanced random-effects model for each data subgroup.

Increased participant experience (learner level) necessitated more observations for reliable assessment: five observations for medical students, seven observations for residents, and eight observations for experienced surgeons. In other words, the assessment tool is better (more reliable) at discriminating performances when the learner is less experienced.

Large differences were found for modality. Specifically, only two observations of cadaveric dissection performances were needed, whereas five observations were needed for VR simulation performances. Similarly, increased fidelity of the graphics in the VR simulator (simulator version 1.3 vs. 2.1 vs. 3.0) necessitated fewer observations (seven, six, and four, respectively).

The structure of training had only a small influence on the number of observations needed: massed practice required five observations compared with the six observations required for distributed practice. Additionally, the simulator-integrated tutoring or greenlighting had no effect on reliability: the number of observations needed with and without the tutor-function during the procedure was 4. Likewise, the same number of observations (five in both cases) would be needed for reliable assessment of participants who had experienced the simulator-integrated tutor-function at any point during their training and for reliable assessment of those who had never experienced it.

Finally, during the initial phase of training when the learning curve is the steepest, (i.e., during the first five procedures), more observations were needed (n = 6) compared with the phase of training when the learning curve starts to plateau and the participant's performance is more stable (which required 4 observations).

Discussion

In this study, we pooled assessment data from over 1,700 unique performances in simulation-based training of a temporal bone surgical procedure in order to analyze the effects of different training variables and learning conditions on assessment reliability.

First, learner experience modified reliability of the assessment: fewer observations of medical students were needed compared with residents and experienced surgeons. One of the attributes of the experienced learner—besides a better performance—is consistency across multiple performances,[30] which results in a low reliability. Many studies on the reliability of technical skills assessment are based on data from novice and expert performances, which, compared to data from more similar groups such as intermediates and experts, can result in overinflating reliability.[31] Altogether, the assessment tool is better at discerning the performances of inexperienced learners, whose performances have larger variability. Our findings align with those of Regehr and colleagues who found that checklist-based assessment of surgical skills is inferior to global rating scales in predicting level of training.[32] A ceiling effect could provide a possible explanation: easier tasks are performed equally well by most participants, resulting in a lower reliability and the need for more observations. This effect further emphasizes that assessment must be targeted to the specific goal of assessment: a test that is good at discriminating between the performances of novices and experts is not necessarily equally useful in discerning between the performances of intermediate trainees. The purpose of assessment should therefore be considered (e.g., is the assessment intended for the longitudinal monitoring of individual learners' progress or for making decisions to certify physicians for practice).

Next, the simulation's fidelity had large effects on reliability: in the most realistic but also most difficult simulation modality, cadaveric dissection, far fewer observations (n = 2) were needed for reliable assessment compared with the 5 necessary in VR simulation. Interestingly, within VR simulation, increasing the graphic fidelity necessitated fewer observations.

In contrast to learner level and fidelity, the structure of training (massed vs. distributed practice) nor training condition (tutoring through greenlighting [whether during the procedure or at some point during training] vs. no greenlighting) had little effect on the reliability of assessment. Our data did not support the hypothesis that simulator-integrated tutoring, a form

of synchronous feedback, makes performances more similar and thereby affects reliability. Since the training in all studies was designed for directed, self-regulated learning, none of the participants received instructor feedback, and the simulator-integrated tutor-function (greenlighting as guidance) was the only source of feedback except for participants' own self-monitoring and self-assessment in relation to the available written instructions. Other types of feedback could potentially affect reliability, but exploring this question would require dedicated studies.

Finally, the initial phase of the learning curve, representing the steepest slope, required more observations than the last plateau phase. This finding highlights one of the problems of surgical skills assessment: on one hand, basing assessment on observations of multiple performances is desirable, but, on the other hand, learning is affected by repeated performance (i.e., participants learn through testing).[33] This learning effect is most marked in the beginning when the improvement between repetitions is largest. The very premise of reliability is that every performance is similar (reflecting true performance) and that the variance between actual performances is due to measurement error. Consequently, reliability is decreased by the learning effect because a major contribution to the difference between performances reflects learning rather than measurement error. This paradox represents an ongoing dilemma in surgical skills assessment and should be considered in determining both the appropriate number of observations and when to assess trainees.

This study highlights some of the psychometric inferences of multiple measurements over time. Assessments in health professions education often require observations of learners in different patient encounters by different types of raters that together yield an unbalanced data structure (each learner receives different numbers of assessments by different assessors, and encounters are at varying levels of difficulty or represent varying patient complexity). Unbalanced data, which are often associated with analyses of workplace-based assessments, are different from "balanced" data (e.g., multiple-choice assessments or objective structured

clinical examinations, which often have standardized sets of encounters). As such, estimates of variance components in unbalanced data can be confounded and may require specialized methods of analysis. In this study of unbalanced data, we used the unbalanced random-effects generalizability study design suggested by Brennan,[27-29] with observations nested in persons crossed with items. We replicated the unbalanced data design for different study subgroups. Given the large-scale data used in our study, we think the resulting variance components and reliability inferences generated could be useful for researchers and practitioners. Additional studies, replicating or similar to ours, can increase educators' understanding of learning effects and the role of measurement precision in assessment. Additionally, such studies may help educators develop or enhance guidelines for analyzing unbalanced data and procuring adequate sample sizes for workplace-based health professions education data.

Other studies on the reliability of simulation-based assessment of surgical technical skills—ranging from the assessment of endovascular surgical expertise to procedures such as flexible optic intubation, knee arthroscopy, and video-assisted thoracoscopic lobectomy— have used G-theory.[34–37] In temporal bone surgery, G-theory has been used to explore the generalizability of a final-product assessment tool (the basis for our modified tool) in temporal bone surgery.[38,39] One study included only a small number of participants and explored the random effects using linear mixed models.[38] The authors attributed 61% of total variance to performance on two bones (bone : resident), reflecting a very large inconsistency in performance across the two bones. However, the raters are potentially confounded in the study design.[38] While the authors present useful implications based on their analysis, many methodologists would agree that such linear mixed models may confound different variance components in unbalanced study designs; and as such, Brennan specifically recommends using the unbalanced random-effects design like the one we used in our study, which yields consistent results.[27] The large sample size that we used in our analysis advances the discussion of factors contributing to variability of performance.

In all these studies using G-theory, reliability analyses are based on data from a single specific training and assessment context. Importantly, other contexts and learning conditions (e.g., other institutions or different assessment modalities) could influence the reliability of the specific assessment tool or simulation-based test due to varying score distributions. Consequently, a G-coefficient cannot be assumed to be a general property of the particular simulation-based assessment of surgical technical skills. Additional studies are needed to examine the effects of different variables on the reliability of assessment in simulation-based surgical skills training. For example, one study reports that reliability of resident performance appraisals was lower in the first year of rating and higher in subsequent years.[8] The authors attributed this difference to the raters adapting to the assessment tool,[8] but the learners adapting to a new learning and assessment context could be another possible explanation. We acknowledge that our study has some limitations. Even though we have explored the effect of different variables on the reliability of assessment using extensive data, our study represents data from a single institution and should be interpreted as such. We used a small group of trained raters, and our data are from studies that took place where learning and practice conditions were extremely controlled. This approach has both advantages and disadvantages: the standardization of training conditions across studies allowed us to explore the specific variables reported, but other potential variables need further exploration. Assessment data from more heterogeneous assessment conditions would potentially allow the study of the effects of multiple institutions and a wider selection of raters with different degrees of rater training on reliability. Our raters received rater training and three of the four raters contributed more than 750 assessments each, reducing the risk of confounding the results with the "rater" learning curve. Finally, some subgroups had only a few performances for us to consider in our analysis since only a few trainees did more than 15 repetitions, the newest version (3.0) of the simulator was only recently released, and recruiting experienced subspecialists is difficult (as they currently have less of a training incentive than other

learners). A strength of our study is the high number of unique performances and total assessments, which allowed us to study the effects of assessment under a range of different training variables and conditions on reliability. Furthermore, participants contributed to only one study each, and recruitment at the level of learner, occurred similarly across studies. All of which reduced the risk of confounding due to contributions to multiple subgroups. Finally, given the varying numbers of observations per trainee in our data, estimates of variance components derived using the unbalanced random-effects design would benefit from further analysis and replication in future studies. Such studies would help inform the stability of the estimates and their confidence intervals, and they could, for example, include specifying whether the assessment item is fixed or consider the learning curve (e.g., modeling growth). Overall, reliability seems less stable than previous skills assessment literature has indicated. Indeed, we found that a number of variables influenced reliability. Our findings emphasize that medical educators should exercise caution when using a specific assessment in novel contexts. Doing so could have implications not only for simulation-based technical skills assessment but also for assessment in medical education more generally. Consequently, reported G-coefficients and D studies should be used only to provide an overall idea of the number of observations needed. For high-stakes assessment, such as certification, reliability must always be carefully studied within in the specific assessment context. We, therefore, recommend the use of G theory to explore reliability of assessment conditions at other institutions, specialties, and assessment contexts.

References

1. Reznick RK. Teaching and testing technical skills. Am J Surg. 1993;165:358–361.
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. Streiner DL, Norman GR, Cairney J. Chapter 8 – Reliability. In: Streiner DL, Norman GR, Cairney J. Health Measurement Scales. 5th ed. Oxford, UK: Oxford University Press; 2015. p. 159–199.
4. Keller LA, Clauser BE, Swanson DB. Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. Adv in Health Sci Educ. 2010;15:717–733.
5. Bilgic E, Watanabe Y, McKendy KM, Ito Y, Vassiliou MC. Reliable assessment of performance in surgery: A practical approach to generalizability theory. J Surg Edu. 2015;72:774–775.
6. Brennan R. Generalizability Theory and Classical Test Theory. Applied Measurement in Education. 2011;24:1–21.
7. Downing SM. Reliability: On the reproducibility of assessment data. Med Educ. 2004;38:1006–1012.
8. 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.
9. Bhatti NI, Cummings CW. Viewpoint: Competency in surgical residency training: Defining and raising the bar. Acad Med. 2007;82:569–573.
10. Carr MM. Program directors' opinions about surgical competency in otolaryngology residents. Laryngoscope. 2005;115:1208–1211.

11. Lui JT, Compton ED, Ryu WHA, Hoy MY. Assessing the role of virtual reality training in Canadian Otolaryngology–Head & Neck Residency Programs: A national survey of program directors and residents. J Otolaryngol Head Neck Surg. 2018;47:61.
12. Frithioff A, Sørensen MS, Andersen SAW. European status on temporal bone training: A questionnaire study. Eur Arch Otorhinolaryngol. 2018;275:357–363.
13. Lui JT, Hoy MY. Evaluating the effect of virtual reality temporal bone simulation on mastoidectomy performance: A meta-analysis. Otolaryngol Head Neck Surg. 2017;156:1018–1024.
14. Sethia R, Kerwin TF, Wiet GJ. Performance assessment for mastoidectomy. Otolaryngol Head Neck Surg. 2017;156:61–69.
15. Andersen SAW, Mikkelsen PT, Sørensen MS. Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance. Laryngoscope. 2019;129:2170–2177.
16. Andersen SAW, Caye-Thomasen P, Sorensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. Laryngoscope. 2015;125:431–435.
17. Andersen SAW, Foghsgaard S, Konge L, Caye-Thomasen P, Sørensen MS. The effect of self-directed virtual reality simulation on dissection training performance in mastoidectomy. Laryngoscope. 2016;126:1883–1888.
18. Andersen SAW, Foghsgaard S, Caye-Thomasen P, Sørensen MS. The effect of a distributed virtual reality simulation training program on dissection mastoidectomy performance. Otol Neurotol. 2018;39:1277–1284.
19. Frendø M, Konge L, Cayé-Thomasen P, Sørensen MS, Andersen SAW. Decentralized virtual reality training of mastoidectomy improves cadaver dissection performance: A prospective, controlled cohort study. Otol Neurotol. 2020;41:476–481.

20. Andersen SAW, Konge L, Caye-Thomasen P, Sørensen MS. Learning curves of virtual mastoidectomy in distributed and massed practice. JAMA Otolaryngol Head Neck Surg. 2015;141:913–918.
21. Andersen SAW, Konge L, Caye-Thomasen P, Sørensen MS. Retention of mastoidectomy skills after virtual reality simulation training. JAMA Otolaryngol Head Neck Surg. 2016;142:635–640.
22. Andersen SAW, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: A randomized controlled trial. Adv Simul (Lond). 2016;1:20.
23. Andersen SAW, Mikkelsen PT, Sørensen MS. The effect of simulator-integrated tutoring for guidance in virtual reality simulation training. Simul Healthc. 2020;15:147–153.
24. Andersen SAW, Guldager M, Mikkelsen PT, Sørensen MS. The effect of structured self-assessment in virtual reality simulation training of mastoidectomy. Eur Arch Otorhinolaryngol. 2019;276:3345–3352.
25. Sørensen MS, Mosegaard J, Trier P. The visible ear simulator: A public PC application for GPU-accelerated haptic 3D simulation of ear surgery based on the visible ear data. Otol Neurotol. 2009;30:484–487.
26. The Visible Ear Simulator. Academic freeware for virtual reality temporal bone surgical training. Available from http://ves.alexandra.dk. [Last accessed June 2, 2020].
27. Brennan RL. Generalizability Theory. New York: Springer; 2001.
28. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. 2014;48:614–622.

29. Zaidi NLB, Kreiter CD, Castaneda PR, et al. Generalizability of Competency Assessment Scores Across and Within Clerkships: How students, assessors, and clerkships matter. Acad Med. 2018;93:1212–1217.

30. Magill RA. The stages of learning. In Magill R~~, Anderson D~~. Motor Learning and Control: Concepts and Applications. New York: McGraw-Hill; 2007. p. 263–289.

31. Hasselager A, Østergaard D, Kristensen T, et al. Assessment of laypersons' paediatric basic life support and foreign body airway obstruction management skills: A validity study. Scand J Trauma Resusc Emerg Med. 2018;26:73.

32. 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.

33. Larsen DP, Butler AC, Roediger HL. Comparative effects of test-enhanced learning and self-explanation on long-term retention. Med Educ. 2013;47:674–682.

34. 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.

35. Graeser K, Konge L, Kristensen MS, Ulrich AG, Hornbech K, Ringsted C. Airway management in a bronchoscopic simulator based setting. Eur J Anaesthesiol. 2014;31:125–130.

36. Jacobsen ME, Andersen MJ, Hansen CO, Konge L. Testing basic competency in knee arthroscopy using a virtual reality simulator: Exploring validity and reliability. J Bone Joint Surg Am. 2015;97:775–781.

37. 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.

38. 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.

39. Butler NN, Wiet GJ. Reliability of the Welling Scale (WS1) for rating temporal bone dissection performance. Laryngoscope. 2007;117:1803–1808.

Figure Legends

Figure 1

Flow chart of the 10 studies (2012 – 2018, Copenhagen, Denmark) of simulation-based assessments of temporal bone surgical performance whose data were included in a Generalizability analysis examining the effect of particular conditions on reliability. The top portion shows the number and level of learners, the number of performances in cadaveric dissection and virtual reality (VR) simulation, and the total number of observations contributed to the overall data pool. The bottom portion lists the information extracted for each observation for use in this study.

Figure 2

Projections in reliability (Decision study) or the effect on the generalizability coefficient of adding more observations regardless of training condition for the overall dataset, which comprises 3,574 simulation-based assessments of temporal bone surgical performance examined in 10 studies from 2012 – 2018 in Copenhagen, Denmark.

Figure 3

Projections in reliability (Decision studies) or the effect on the generalizability coefficient of adding more observations for different training variables: (A): Learner level; (B): Training modality (dissection of human cadaveric temporal bones vs. virtual reality [VR] simulation); (C) Simulator version (iterative increase in the fidelity of the graphics across versions 1.3, 2.1, and 3.0); (D) Training structure (massed practice vs distributed practice); (E) Tutoring during the procedure (the simulator-integrated tutor function on vs. off during the specific procedure); (F) Tutoring during training (the simulator-integrated tutor function used for at least one procedure vs. never during the training program); and (G) Slope of the learning curve (grouped by the number of procedures performed). The authors used data from 10 studies with simulation-based assessments of temporal bone surgical performance from 2012 – 2018 in Copenhagen, Denmark.

Table 1

Overview of the Distribution of 1,723 Unique Performances of a Procedure[a] Among Participants, 2012-2018

Subgroups	No. (%[b]) of performances	Estimated marginal means of performances, sum score (95% confidence interval)
Raters		
Rated by Rater 1	1,721 (48.2[c])	15.8 (15.6–16.1)
Rated by Rater 2	772 (21.6[c])	14.3 (14.0–14.7)
Rated by Rater 3	117 (3.3[c])	15.1 (14.4–16.0)
Rated by Rater 4	964 (27.0[c])	18.1 (17.8–18.4)
Rated by 1 rater	9 (0.5)	N/A
Rated by 2 raters	1,577 (91.5)	N/A
Rated by 3 raters	137 (8.0)	N/A
Learner levels		
Experienced otosurgeons	32 (1.9)	19.7 (18.7–20.7)
Residents	498 (28.9)	15.4 (15.1–15.8)
Medical students	1,193 (69.2)	16.2 (15.8–16.5)
Training modality		
VR simulation	1,592 (92.4)	16.2 (16.0–16.5)
Cadaveric dissection	131 (7.6)	13.0 (12.4–13.6)
VR simulator fidelity[d]		
Version 1.3	642 (40.3[e])	15.5 (15.1–15.9)
Version 2.1	904 (56.8[e])	17.7 (17.3–18.1)
Version 3.0	46 (2.9[e])	18.3 (17.4–19.2)
Training structure (VR simulation)		
Distributed practice	1,112 (69.8[e])	17.1 (16.7–17.5)
Massed practice	480 (30.2[e])	15.7 (15.3–16.1)
Tutoring during procedure (VR simulation)		
With tutoring during the procedure	377 (23.7[e])	17.6 (17.1–18.0)
Without tutoring during the procedure	1,215 (76.3[e])	15.8 (15.4–16.2)
Tutoring during training (VR simulation)		
Tutored cohort	878 (55.2[e])	16.3 (16.0–16.7)
Non-tutored cohort	714 (44.8[e])	16.8 (16.4–17.2)
Part of the learning curve		
Initial phase (procedures #1–5)	856 (49.7)	14.4 (14.2–14.7)
Interim phase (procedures #6–10)	444 (25.8)	16.0 (15.6–16.3)
Plateau phase (procedures #11–15)	398 (23.1)	16.0 (15.7–16.4)
Post-plateau phase (procedures #16 and beyond)	25 (1.5)	16.0 (14.9–17.1)

Abbreviations: N/A, not applicable; VR, virtual reality.
[a]The principal temporal bone procedure: complete anatomical mastoidectomy ± posterior tympanotomy (the further drilling of the small bony plate between the facial nerve and chorda tympani).
[b]Unless otherwise indicated, percentage of 1,723 performances.
[c]Percentage of 3,574 assessments.
[d]Each successive version had improved fidelity.
[e]Percentage of 1,592 VR performances.

Table 2

Variance Components for a Generalizability Study Examining Factors in Assessment of Surgical Technical Skills[a]

Effects	*df*	VC	%VC	Explanation
person	245	0.012	4.8	How well the tool discriminates high and low performers (true variance)
observation : person	3,328	0.011	4.6	Variability in observations (rater effect)
item	25	0.090	36.9	Variability in item difficulty
person x item	6,125	0.026	10.8	Interaction between learners and items (item specificity)
residual error	83,200	0.105	43.0	Unexplained variance

Abbreviations: *df*, degrees of freedom; VC, variance component.

[a]The authors used an unbalanced random-effects design: [(observation : person) x item]

Figure 1

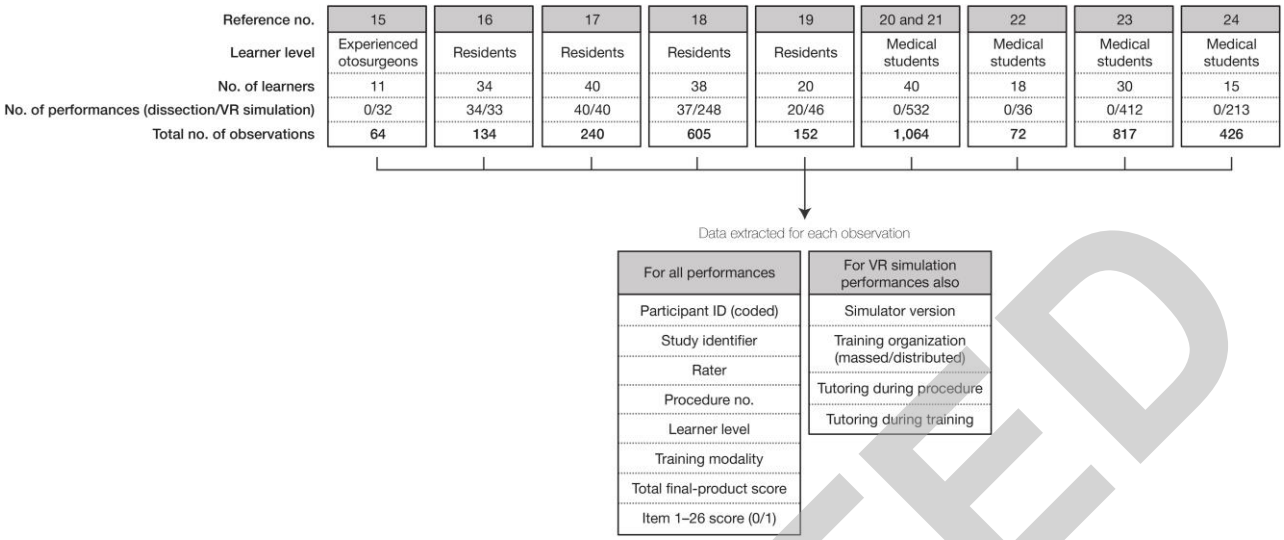

Figure 2

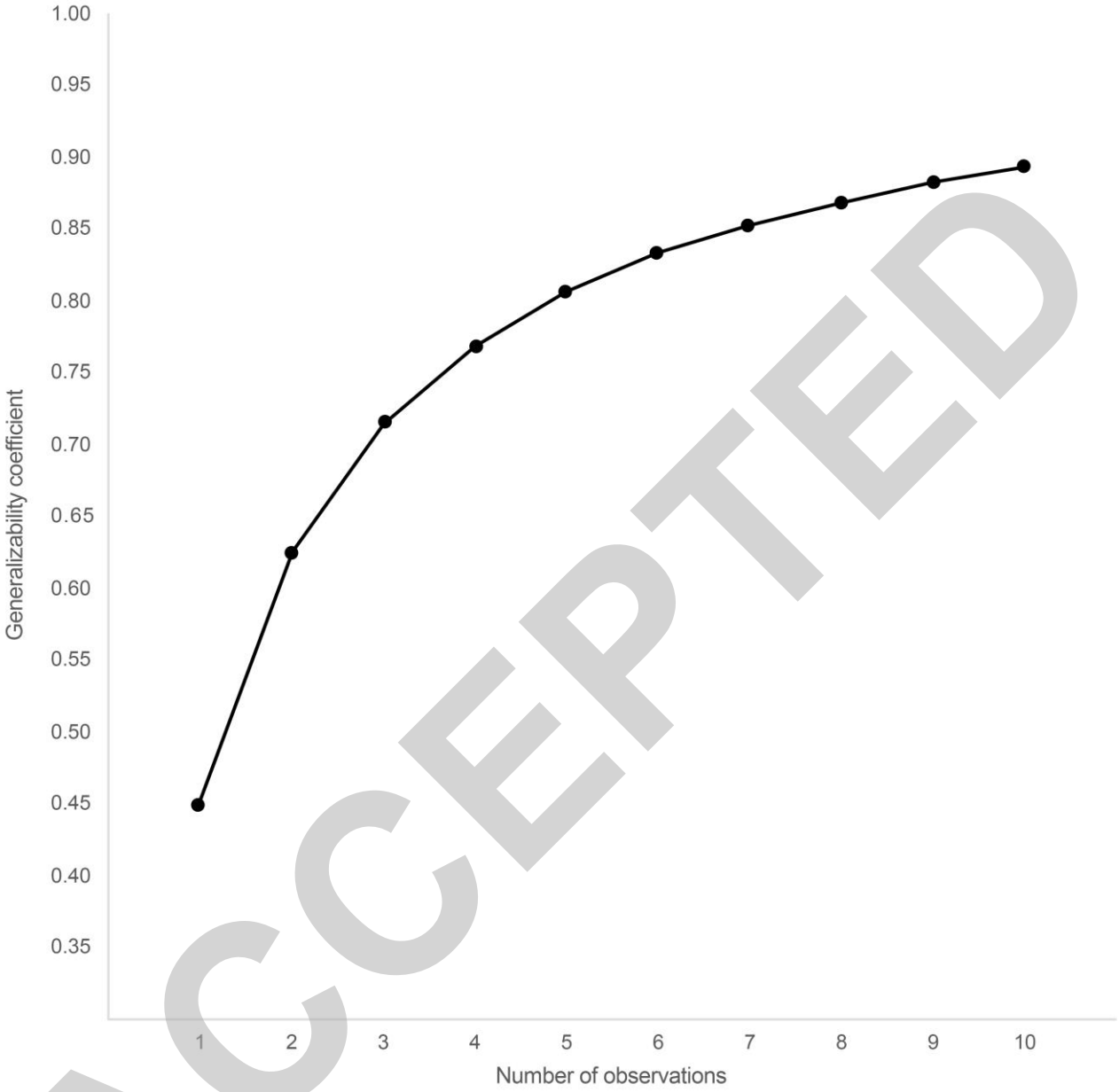

Figure 3