Number of Operative Performance Ratings Needed to Reliably Assess the Difficulty of Surgical Procedures

Kenneth L. Abbott, MS,[*] Xilin Chen, MPH,[†] Michael Clark, PhD,[‡] Nikki L. Bibler Zaidi, PhD,[*] David B. Swanson, PhD,[§,¶] and Brian C. George, MD, MAEd[†]

[*]University of Michigan Medical School, Ann Arbor, Michigan; [†]Center for Surgical Training and Research, Department of Surgery, University of Michigan, Ann Arbor, Michigan; [‡]Consulting for Statistics, Computing, and Analytics Research, University of Michigan, Ann Arbor, Michigan; [§]American Board of Medical Specialties, Chicago, Illinois; and [¶]University of Melbourne Medical School, Melbourne, Victoria

OBJECTIVE: The profession of surgery is entering a new era of "big data," where analyses of longitudinal trainee assessment data will be used to inform ongoing efforts to improve surgical education. Given the high-stakes implications of these types of analyses, researchers must define the conditions under which estimates derived from these large datasets remain valid. With this study, we determine the number of assessments of residents' performances needed to reliably assess the difficulty of "Core" surgical procedures.

DESIGN: Using the SIMPL smartphone application from the Procedural Learning and Safety Collaborative, 402 attending surgeons directly observed and provided workplace-based assessments for 488 categorical residents after 5259 performances of 87 Core surgical procedures performed at 14 institutions. We used these faculty ratings to construct a linear mixed model with resident performance as the outcome variable and multiple predictors including, most significantly, the operative procedure as a random effect. We interpreted the variance in performance ratings attributable to the procedure, after controlling for other variables, as the "difficulty" of performing the procedure. We conducted a generalizability analysis and decision study to estimate the number of SIMPL performance ratings needed to reliably estimate the difficulty of a typical Core procedure.

RESULTS: Twenty-four faculty ratings of resident operative performance were necessary to reliably estimate the difficulty of a typical Core surgical procedure (mean dependability coefficient 0.80, 95% confidence interval 0.73-0.87).

CONCLUSIONS: At least 24 operative performance ratings are required to reliably estimate the difficulty of a typical Core surgical procedure. Future research using performance ratings to establish procedure difficulty should include adequate numbers of ratings given the high-stakes implications of those results for curriculum design and policy. (J Surg Ed 76:e189–e192. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Core, procedure, difficulty, performance, generalizability, dependability

COMPETENCIES: Medical Knowledge, Practice-Based Learning and Improvement, Systems-Based Practice

INTRODUCTION

Nationwide surveys have identified widespread concern about the readiness of general surgery residents for entry into independent practice or fellowship.[1-3] Operative experience for many surgical procedures may be inadequate for the development of competency, including some procedures deemed essential or "Core"[4] to the profession of surgery.[1,5-7] One recent study found that residents performing a typical Core procedure during the last 6 months of their residency program were rated by attending surgeons as being ready for independent practice only 80% of the time.[6] This evidence suggests that some newly minted surgeons may be unable to independently and competently perform the full breadth of Core general surgery procedures.

Funding sources: This study was funded by a grant from the American Board of Surgery and a Graduate Medical Education Innovations Grant from the University of Michigan Medical School. The initial development of SIMPL was funded via grants from Massachusetts General Hospital, Northwestern University, and Indiana University. Subsequent development has been funded by contributions from the members of the Procedural Learning and Safety Collaborative (PLSC, http://www.procedurallearning.org).

Correspondence: Inquiries to Brian C. George, MD, MAEd, Center for Surgical Training and Research, Department of Surgery, University of Michigan, Ann Arbor, MI; e-mail: bcgeorge@med.umich.edu

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2019.07.008

Fortunately, there are emerging methods and data sources that may be used to address these issues. For example, "big data" in the form of longitudinal trainee assessments can be used to examine how difficult it is to learn to perform individual procedures. Such studies would have high-stakes implications for curriculum redesign and policy. Unfortunately, the methods for estimating the relative difficulty of individual procedures have not been fully developed. Specifically, it remains unknown how many observations of individual operative performances are needed to reliably estimate the difficulty of "Core"[4] surgical procedures. This study aims to address that methodological gap.

METHODS

For this study, we used longitudinal data from a registry managed by the Procedural Learning and Safety Collaborative (PLSC), a nonprofit research consortium with member institutions across the United States; these records include faculty ratings of intraoperative resident performance and autonomy for a longitudinal sample of each resident's operative experiences. Faculty completed these ratings within 72 hours of each case via the SIMPL[9,10] smartphone application. Possible SIMPL performance ratings include *unprepared/critical deficiency*, *inexperienced with procedure*, *intermediate performance*, *practice-ready performance*, or *exceptional performance*. We excluded cases involving multiple procedures, multiple residents, or noncategorical surgical residents in order to simplify the analysis and better illuminate effects of interest. Ultimately, we analyzed 5259 workplace-based assessments (WBAs) from 14 institutions, with 402 attending surgeons observing 488 categorical residents performing 87 Core surgical procedures between September 2015 and January 2017. The dataset was unbalanced, meaning that not every resident would be rated by every faculty member on the performance of every procedure.

We used these performance ratings to construct a linear mixed model; this choice of model allowed us to examine multiple sources of variance without having to drop observations from our dataset in pursuit of a balanced design. Our mixed model used resident performance as the outcome variable and multiple predictors including, most significantly, operative procedure as a random effect. We selected other random effects (program, rater, and resident) and fixed effects (postgraduate year, week within academic year, and patient-related complexity) based upon prior research establishing relationships between these variables and SIMPL performance ratings.[6] After controlling for these variables, we interpreted the variance in the performance ratings attributable to the procedure as the "difficulty" of performing the procedure.

We used generalizability (G) theory[11] to estimate the reliability of estimates of the difficulty of a typical Core surgical procedure as a function of the number of observations (see Supplement). G theory is often used to identify testing conditions necessary to achieve a desired level of score reliability for multifaceted assessments. A decision (D) study determines the number of ratings needed to achieve a suitably high dependability coefficient—usually at least 0.8.[12] We completed a D study to estimate the number of SIMPL performance ratings needed to reliably estimate the difficulty of a typical Core procedure—i.e., the theoretical Core procedure representing the amalgam of several dozen Core procedures in our data set, weighted according to procedure frequency.

RESULTS

There were only 2 ratings for *unprepared/critical deficiency*, so we collapsed *unprepared/critical deficiency*, and *inexperienced with procedure* into a single category. On the resulting scale from 1 to 4, the mean (SD) SIMPL performance rating was 2.4 (0.7). Table 1 provides descriptive statistics and variance components from the linear mixed model. Cells in the standard deviation (SD) column can be interpreted as an index of the "true" (i.e., if no measurement error were present) variability associated with each row, expressed on the scale on which ratings were provided. For example, the SD of

TABLE 1. SIMPL Operative Performance Rating Descriptive Statistics and Variance Components from Linear Mixed Model

Source of Variance	Sample Size	Median # of Observations	Variance Component	Standard Deviation	Percentage of Variance	Interpretation
Procedures	87	23	0.07	0.26	15%	Variability due to procedure difficulty
Programs	14	238.5	0.00	0.05	0%	Variability due to program-level factors
Faculty raters	402	5	0.11	0.32	23%	Variability due to rater stringency
Residents	488	6	0.05	0.22	10%	Variability due to resident proficiency
Residual			0.24	0.49	52%	

0.26 for procedures indicates that 95% of procedure difficulties fell between plus or minus twice that SD. The variability in rater stringencies was somewhat larger than this, and the variability in overall resident proficiency was somewhat smaller. The D study suggested that at least 24 observations of operative performance, across all residents instead of per resident, were necessary to accurately estimate the difficulty of a typical Core surgical procedure (mean dependability coefficient 0.80, 95% confidence interval 0.73-0.87).

DISCUSSION

All performance assessments are subject to multiple sources of variability, even when using WBAs such as SIMPL that follow best practices.[13] Fortunately, G theory allows us to investigate this variability and establish how many ratings are needed to achieve reproducibility. We found that 24 WBA ratings of operative performance were sufficient to reliably estimate the difficulty of a typical Core surgical procedure. This is lower than the 40 ratings needed to reliably estimate an individual surgical trainee's performance, according to a recent study.[8] Our findings can be used as guidance on how to reproducibly achieve what one recent paper referred to as "shared subjectivity"[14]—in this case, a shared subjective view of how difficult it is to competently perform a procedure.

LIMITATIONS

This study has several limitations. First, no ratings were available for 40 Core surgical procedures, presumably because they were rarely performed by residents. It is, therefore, unclear how well these results generalize to procedures not included in the dataset, and future studies with larger datasets are needed to investigate those procedures.

Second, nonrandom samples of attendings and residents at the 14 included programs may have chosen to participate, and the results may have been influenced by residents and faculty for whom larger numbers of ratings were available. Hence, estimates of the number of ratings needed to estimate procedure difficulty may not apply to other study settings.

Third, our projections focused on a hypothetical "typical" Core procedure; individual Core procedures may require more or less than 24 ratings for reliable estimation of difficulty. Research using larger datasets would allow for the estimation of the difficulty of individual procedures.

Fourth, the results indicated that residents' skills have less impact on ratings than differences in faculty stringency (i.e., hawk or dove effects).[15] This rater-level variation increases the number of faculty ratings needed to estimate procedure difficulty, and it seems possible that additional rater training aimed at establishing a greater shared understanding of performance benchmarks might reduce the number of ratings needed to estimate procedure difficulty.

Fifth, there may have been situations in which procedure difficulty was confounded with one or more other factors. For example, suppose a set of procedures are all difficult, and that faculty raters for these procedures are all stringent in their evaluations of resident performance; in such circumstances, procedure difficulty would be confounded with rater stringency. We attempted to separate the effects of all these variables via mixed modeling, and we expect the risk of confounding to diminish with increasing numbers of observations, but we speculate that, for some rarer procedures, difficulty might have been confounded with one or more other variables.

Finally, estimates of procedure difficulty may reflect both the inherent technical difficulty of performing the procedure and the difficulty of learning to perform the procedure. Ratings of resident performance in the study dataset may have occurred relatively early or late in a resident's experience with a procedure. Ratings gathered early in a resident's experience with a procedure seem likely to make the procedure look technically harder; the reverse is true if ratings were gathered after a resident had substantial experience with the procedure. Ideally, the analysis would consider when ratings were obtained in a resident's "learning curve" for a procedure, but residents' prior experience with a procedure was not available for use in this analysis. This might result in fewer ratings being required to estimate a procedure's technical difficulty and may also make it possible to estimate the "learning trajectory" for frequently performed procedures; the latter seems highly relevant to designing residents' training experience to ensure mastery of Core procedures. At the same time, because the mixed model used in our analysis included both program year and week within the program year, the time required to learn the procedure was, to a degree, controlled for.

Despite these limitations, study results should provide surgical educators and researchers with useful guidance until larger studies can be conducted.

FUTURE RESEARCH

Guided by these results, future research using performance ratings to establish procedure difficulty should include numbers of ratings that are sufficient to ensure reliability of data because the results of these analyses may have high-stakes implications for curriculum design and policy. For example, residency programs may need to provide additional instruction for those Core general surgery procedures that are

most difficult to learn. Should volume for certain Core procedures be low at a given institution, it may be desirable for residency programs to supplement instruction with simulation. Alternatively, analyses of procedure difficulty may be used to reset expectations for what can feasibly be learned during a residency program of fixed duration.

CONCLUSIONS

At least 24 operative performance ratings are required to reliably estimate the difficulty of a typical Core general surgical procedure. Future research using performance ratings to establish procedure difficulty should include adequate numbers of ratings given the high-stakes implications of those results for curriculum design and policy.

ACKNOWLEDGMENTS

The authors thank the program directors, coordinators, faculty, and residents from the participating PLSC programs for making this study possible. They also wish to thank all the institutional members of PLSC for their help in designing, developing, and supporting SIMPL.

REFERENCES

1. Bell RH Jr., Biester TW, Tabuenca A, et al. Operative experience of residents in US general surgery programs: a gap between expectation and experience. *Ann Surg*. 2009;249:719–724.
2. Mattar SG, Alseidi AA, Jones DB, et al. General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Ann Surg*. 2013;258:440–449.
3. Napolitano LM, Savarise M, Paramo JC, et al. Are general surgery residents ready to practice? A survey of the American College of Surgeons Board of Governors and Young Fellows Association. *J Am Coll Surg*. 2014;218:1063–1072. e1031.
4. Resident Education SC. Curriculum outline for general surgery 2017-2018. In: 2017.
5. Malangoni MA, Biester TW, Jones AT, Klingensmith ME, Lewis FR Jr. Operative experience of surgery residents: trends and challenges. *J Surg Educ*. 2013;70:783–788.
6. George BC, Bohnen JD, Williams RG, et al. Readiness of US general surgery residents for independent practice. *Ann Surg*. 2017;266:582–594.
7. Strosberg DS, Quinn KM, Abdel-Misih SR, Harzman AE. Redefining the Surgical Council of Resident Education (SCORE) curriculum: a comparison with the operative experiences of graduated general surgical residents. *Am Surg*. 2018;84:526–530.
8. Williams RG, Swanson DB, Fryer JP, et al. How many observations are needed to assess a surgical trainee's state of operative competency. *Ann Surg*. 2017;269:377–382.
9. Bohnen JD, George BC, Williams RG, et al. The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): early experience from a multi-institutional trial. *J Surg Educ*. 2016;73:e118–e130.
10. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ*. 2014;71:e90–e96.
11. Brennan R. Statistics for Social Science and Public Policy: Generalizability Theory. New York: Springer; 2001.
12. McAleer S, Chandratilake M. Choosing instruments for assessment. Walsh K, editor. Oxford Textbook of Medical Education, Oxford: Oxford University Press; 2013:432–440.
13. Williams RG, Kim MJ, Dunnington GL. Practice guidelines for operative performance assessments. *Ann Surg*. 2016;264:934–948.
14. ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2018;94:333–337.
15. Douglas C. A day in the country. *Bmj*. Vol 328. Copyright © 2004. BMJ Publishing Group Ltd.; 2004. p. 1573.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2019.07.008.