The American Journal of Surgery (2016) 211, 426-430

The American Journal of Surgery®

Association for Surgical Education

Reliable assessment of operative performance

Elif Bilgic, B.Sc.[a], Yusuke Watanabe, M.D.[a,b], Katherine McKendy, M.D.[a], Amani Munshi, M.D.[a], Yoichi M. Ito, Ph.D.[c], Gerald M. Fried, M.D.[a], Liane S. Feldman, M.D.[a], Melina C. Vassiliou, M.D., M.Ed[a,*]

[a]*Department of Surgery, Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, 1650 Cedar Avenue, D6-257, Montreal, QC, H3G1A4, Canada;* [b]*Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan; and* [c]*Department of Biostatistics, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan*

KEYWORDS:
Generalizability theory;
Reliability;
Surgery;
Assessment

Abstract

BACKGROUND: There is no consensus regarding the number of intraoperative assessments required to reliably measure trainee performance. This study used generalizability theory (GT) to describe factors contributing to score variance and to estimate the number of assessments needed to achieve high standards of reliability.

METHODS: While performing laparoscopic procedures, trainees were assessed by the attending surgeon using Global Operative Assessment of Laparoscopic Skills (GOALS). Data were collected prospectively (2-month intervals), assessing each trainee multiple times. Reliability coefficient was calculated using trainees, cases, and raters as factors.

RESULTS: Eighteen trainees were included for a total of 65 assessments. Total variance in scores was accounted for as follows: 66.1% by trainees, 31.6% by the interaction between trainees and cases, and 2.3% by raters. At least 3 cases are required for reliable scores using GOALS.

CONCLUSIONS: Trainees accounted for most of the variance in GOALS scores with a minimum of 3 cases required to improve the reliability of the scores obtained. These data may guide the implementation of performance assessments in surgical training programs.

The way we train and assess surgeons has been evolving from case numbers and in-training evaluations to direct observations of performance using workplace-based assessments. Various tools and instruments to measure operative performance are available to document that surgical trainees have achieved proficiency in a certain task or procedure.[1] The General Surgery Milestone Committee has recommended regular operative performance assessments as milestones for all general surgery residents.[2] However, their practical application in residency programs is still not well established, and for most of them, little evidence is available on how to make decisions based on scores obtained using these tools. In order for these metrics to be used for trainee assessment, they must be reliable, that is, the score should be consistent when the same trainee is assessed under the same conditions (assuming that the trainee's

The study had no sponsors or involvement with sponsors.

* Corresponding author. Tel.: +1-514-934-1934x44330; fax: 514-934-8210.

E-mail address: melina.vassiliou@mcgill.ca

Manuscript received June 16, 2015; revised manuscript September 25, 2015

0002-9610/$ - see front matter © 2016 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amjsurg.2015.10.008

skills are stable). It is essentially impossible, however, to create the same conditions in the operating room from one case to another. Apart from trainee skill, scores may be affected by whether there is an easy or hard rater, the difficulty of the particular case, and the type of procedure. Furthermore, these factors may all be interacting simultaneously and can significantly impact scores and put the reliability of a single evaluation into question, especially, if the score is going to be used to make decisions about promotion or remediation of trainees.

Inter-rater reliability (raters), test-retest reliability (cases), and other so-called "classic" methods are commonly used to assess reliability. Even though these are very useful, they have some limitations because the impact of raters or cases on scores is evaluated separately, and interactions between these factors cannot be taken into account. Generalizability theory (GT) is a statistical method in which the different factors contributing to variations in assessment scores and their interactions are taken into account when estimating reliability. GT permits the integration of multiple factors that might simultaneously impact a trainee's score into one reliability coefficient.[3] Furthermore, for a given assessment tool, once the overall reliability is calculated using GT, the number of raters or cases needed to reliably measure a trainee's skill level can be estimated.[4]

The Global Operative Assessment of Laparoscopic Skills (GOALS) was developed to measure basic, generic laparoscopic skills and has been used to evaluate residents by direct observation in the operating room in multiple different studies and under various conditions.[5,6] The initial publications on GOALS reported excellent inter-rater reliability for different raters assessing residents removing the gallbladder from the liver bed.[7] No study to date, however, has used GT to assess the reliability of GOALS scores obtained for trainees performing various procedures and being assessed by various raters.

The purposes of this study were to apply GT: (1) to examine the impact of trainees, cases, and raters on assessment scores using GOALS; (2) to determine the reliability coefficient of GOALS scores for one assessment by one rater; and (3) to evaluate the number of cases needed to obtain reliable GOALS scores.

Methods

Setting

This prospective study was conducted from July 2014 to January 2015 and was approved by the local Ethics Review Board of McGill University. General surgery residents at all levels and fellows were included. Using GOALS, trainees were assessed by the attending surgeon after each case. Trainees were assessed on multiple occasions within a 2-month interval.

Instrument

The GOALS is an assessment tool designed to measure basic laparoscopic skills and is reported in detail elsewhere.[7] Briefly, it includes 5 domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. Each domain is scored in a 5-point Likert scale with descriptive anchors at 1, 3, and 5. Scores range from 5 to 25. There is evidence supporting the validity of GOALS as a measure of generic laparoscopic skills when used for direct observation in the operating room. It has been used to measure skills in different institutions and over a wide range of both basic and advanced laparoscopic procedures.[5,8]

Rating process

No additional training on how to use GOALS was provided to the attending surgeons; however, all of them had experience using the tool in the past to assess resident performance in the operating room. The primary investigator was present in the operating room to provide the attending surgeons with a paper copy of the GOALS assessment tool at the end of every case. Attending surgeons were asked to complete the assessment immediately after the case. Assessments were only accepted if they were completed on the same day of the procedure.

Statistical analysis

GT was used to determine the impact of the factors on assessment scores and the overall reliability coefficient for the total GOALS score. Decision study was then applied to determine the number of cases needed to reliably assess a trainee's skill level using GOALS.[9] Using JMP, version 11 (SAS Institute Inc, Cary, NC), the variance of each component and the impact of each factor on assessment scores were calculated using analysis of variance, based on an unbalanced data set that was collected. The generalizability coefficient (overall reliability coefficient) and the number of cases required were also calculated. Trainees (t), cases (c), and raters (r) were included as factors along with their interaction terms (fully nested design).[10] Cases and raters were labeled as random. The number of cases needed to achieve the recommended standards of a minimum reliability of .8 was determined.[11]

Results

Eighteen trainees (3 Post Graduate Year (PGY)2, 1 PGY3, 3 PGY4, 8 PGY5, and 3 fellows) underwent a median of 3 GOALS assessments (Interquartile range, 2 to 5) each (total of 65 assessments) by 9 attending surgeons. Ten raters participated in the study; however, one rater assigned almost perfect scores to all residents and was therefore excluded from the analysis. The laparoscopic procedures