The American Journal of Surgery 213 (2017) 273–276

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.ajconline.org

Association for Surgical Education

Psychometric properties of the Global Operative Assessment of Laparoscopic Skills (GOALS) using item response theory☆

Yusuke Watanabe [a, b], Amin Madani [a], Yoichi M. Ito [c], Elif Bilgic [a], Katherine M. McKendy [a], Liane S. Feldman [a], Gerald M. Fried [a], Melina C. Vassiliou [a, *]

[a] Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, 1650, Ave. Cedar, Montreal, QC, H3G 1A4, Canada
[b] Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Kita 15, Nishi 7, Kita-ku, Sapporo, 060-8638, Japan
[c] Department of Biostatistics, Hokkaido University Graduate School of Medicine, Kita 15, Nishi 7, Kita-ku, Sapporo, 060-8638, Japan

ARTICLE INFO

Article history:
Received 8 May 2016
Received in revised form
19 September 2016
Accepted 24 September 2016

Keywords:
Item response theory
Clinical competence
Performance assessment
Workplace-based assessment
Psychometric property
Surgical education

ABSTRACT

Background: The extent to which each item assessed using the Global Operative Assessment of Laparoscopic Skills (GOALS) contributes to the total score remains unknown. The purpose of this study was to evaluate the level of difficulty and discriminative ability of each of the 5 GOALS items using item response theory (IRT).

Methods: A total of 396 GOALS assessments for a variety of laparoscopic procedures over a 12-year time period were included. Threshold parameters of item difficulty and discrimination power were estimated for each item using IRT.

Results: The higher slope parameters seen with "bimanual dexterity" and "efficiency" are indicative of greater discriminative ability than "depth perception", "tissue handling", and "autonomy".

Conclusions: IRT psychometric analysis indicates that the 5 GOALS items do not demonstrate uniform difficulty and discriminative power, suggesting that they should not be scored equally. "Bimanual dexterity" and "efficiency" seem to have stronger discrimination. Weighted scores based on these findings could improve the accuracy of assessing individual laparoscopic skills.

1. Introduction

While assessments of surgical performance are crucial in surgical training, their accuracy remains a challenge. In order to successfully assess operative performance using global rating scale (GRS)–based assessments, they should precisely estimate the operative ability (latent trait) and discriminate between surgeons who perform better or worse.

When using GRS-based assessments such as the Global Operative Assessment of Laparoscopic Skills (GOALS), total scores (entire performance scores) are commonly used by calculating either a simple mean score or the sum score of each equally weighted item.[1] Previous studies have demonstrated the internal consistency of GOALS based on an analysis that supports certain psychometric properties of the tool,[2] and assuming that each item has the same discriminative ability. However, the contributions of each item being assessed by GOALS to entire performance scores and the magnitude of the differences between scores for each item are not known. The magnitude of the difference between two scores within an item may also be inconsistent and the differences in observed behaviors between scores may not be the same from one item to the other.

Latent contributions of each item to final scores and the discriminative ability of each item may vary with each performance assessment, and this could have various impacts on the entire performance score. For instance, GOALS consists of 5 items: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy, scored using a 5-point Likert scale to assess laparoscopic operative performance.[2] Interpretation of GOALS is based on the total score, which is calculated by simply summing the scores for each of the equally weighted items (score range 5–25). Obtaining a

☆ Association for Surgical Education Paper Session (podium presentation), Association for Surgical Education (ASE) 2015 Annual Meeting, April 23–25, 2015, Chicago, IL.

* Corresponding author. McGill University Health Centre, 1650 Cedar Avenue, Room L9-313, Montreal, QC, H3G 1A4, Canada.
E-mail address: melina.vassiliou@mcgill.ca (M.C. Vassiliou).

http://dx.doi.org/10.1016/j.amjsurg.2016.09.050
0002-9610/

score of 4 for depth perception may or may not be as hard as getting a score of 4 for efficiency. Furthermore, the psychometric difference between a score of 1 and 2 may not be the same as the difference between a score of 4 and 5 for the same item.

When interpreting assessment scores, the significance of these differences between items could be a potential source of measurement error and might be taken into account in order to improve the accuracy of the measurement tool. Traditionally, educators have used inter-rater reliability, test retest reliability, item-totals or item-item correlations (including internal consistency, i.e., Cronbach alpha) to evaluate the internal structure of an assessment tool, collectively known as "classical test theory (CTT)".[3,4] There are concerns that these methods have limitations due to the lack of generalizability about assumptions of item equivalence.[4–6] IRT was established as a method to overcome the potential biases associated with CTT. The purpose of this study was to determine the item difficulty and discriminative power of each of the 5 GOALS items and to calculate calibration scores using item response theory (IRT).

2. Methods

A sample of general surgery residents at a single teaching institution underwent GOALS assessments by attending surgeons. The prospectively collected GOALS assessments (between 2003 and 2015) were completed for a range of different laparoscopic procedures. These assessments of surgical trainees were completed by the attending surgeon responsible for each case. All of the attending surgeons in this study were familiar with assessing operative skills using GOALS. Approval from the McGill University Institutional Review Board was obtained, and consent was obtained from all of the participants.

2.1. Global Operative Assessment of Laparoscopic Skills (GOALS)

GOALS is a tool used to evaluate generic laparoscopic skills with five domains (depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy), each with an ordinal response format (score 1 to 5).[2] Total score is calculated as the sum of all five items, such that all items are given equal weight and therefore, the total score ranges from 5 to 25. A higher score indicates better laparoscopic operative performance.

2.2. Data analysis and statistical methods

Total GOALS scores were calculated by the sum of the scores for each of the 5 equally weighted GOALS items (score range 5–25). Each of the 5 items of the tool was analyzed using the polytomous IRT model (graded response model), reflecting the ordinal nature of item scores (corresponding to each of the 5-point Likert scale levels).[7] This model is an extension of the basic IRT models including the one-parameter model such as Rash model and the two-parameter model. Prior to applying IRT modelds, unidimensionality and local independence should be met as basic assumptions. For investigating unidimensionality assumption, the polychoric correlations between items were calculated. Local independence was assessed by estimating the expectation of scores of each item conditional on subject's ability and item parameters. The residuals can be estimated from subtraction of observed scores from expected scores.

In item response theory, a subject's ability is expressed as "latent trait: θ" which is reported on a spectrum typically from -4 to $+4$, where 0 is the average score and the more positive the value, the greater level of difficulty is required to achieve this increment in score. Threshold parameters indicate item difficulty levels and slope parameters reflect item discrimination ability. Item difficulty means how much of the trait (operative skill) a subject needs to have to obtain a certain score on the item. For item difficulty parameters, four threshold parameters for each item were estimated. Lower values in the threshold parameters denote easier skills being measured on each scale of an item. The slope parameter (discrimination parameter) is a measure of the differential capability of an item. Higher slope parameter values indicate items that have a greater ability to differentiate skill levels of subjects. IRT calibrated GOALS scores were calculated based on weighted scores from threshold parameters and slope parameters. All analyses were performed using JMP version 11 (SAS Institute Inc, Cary, NC) or SAS version 9.4 (SAS Institute Inc, Cary, NC).

3. Results

A total of 396 GOALS assessments are included in the analysis (Table 1). The laparoscopic procedures mostly included cholecystectomies, abdominal wall hernia repairs (inguinal, ventral, and incisional hernia repairs), and colorectal surgery. The range of polychoric correlations were from 0.655 to 0.850 and the cumulative proportion of the first eigenvalue of the polychoric correlation matrix was 0.811. The range of correlations were from -0.443 to -0.126, indicating the correlations of residuals between items were low. Thus, the assumptions of unidimensionality and local independence was met.

The total scores ranged from 8 to 25. Mean scores (standard deviation: SD) of each item and the total scores are provided in Table 2. The threshold parameters and the slope parameter for each of the 5 GOALS items are shown in Table 3. For the difficulty parameters, "Bimanual dexterity", "efficiency" and "autonomy" are more positive, indicating that it is more difficult to be graded a higher score compared to "depth perception" and "tissue handling". For the slope (discrimination) parameters, "bimanual dexterity" and "efficiency" had higher parameters, which are an indication of greater discriminative ability than "depth perception", "tissue handling", and "autonomy". Fig. 1 illustrates the relationship between IRT calibrated scores and original scores (simply equally weighted sum scores). For subjects who had the same original GOALS scores such as 15, IRT calibrated GOALS scores varied from -1.45 to -0.70.

4. Discussion

This study demonstrates further evidence for the psychometric properties of the GOALS assessment, including the magnitude of the differences between the same scores across items and between two scores (e.g., 1 to 2, 4 to 5) within an item. The GOALS assessment fulfilled the preconditions including unidimentionality and local

Table 1
Characteristics of the GOALS assessments (n = 396). Results are presented as n (%).

	N (%)
Training level	
PGY1/2	27 (7)/49 (13)
PFY 3/4/5	40 (10)/75 (19)/112 (29)
Fellow/attending surgeon	72 (18)/15 (2)
Types of laparoscopic procedure	
Cholecystectomy	173 (44)
Ventral Hernia	63 (16)
LIHR	57 (14)
Colorectal surgery	49 (12)
Upper GI	18 (5)
Miscellaneous[a]	36 (9)

PGY, post-graduate year; LIHR, Laparoscopic Inguinal Hernia Repair; GI, gastrointestinal.
[a] Miscellaneous includes appendectomies, solid organs.