Identifying the Essential Portions of the Skill Acquisition Process Using Item Response Theory

Saseem Poudel[*,†], Yusuke Watanabe[*,‡], Yo Kurashima, MD, PhD, [*,§] Yoichi M. Ito[‖], Yoshihiro Murakami[¶], Kimitaka Tanaka[*], Hiroshi Kawase[*,#], Toshiaki Shichinohe[*] and, Satoshi Hirano[*]

[*]Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Japan; [†]Department of Surgery, Steel Memorial Muroran Hospital, Muroran, Japan; [‡]Department of General Surgery, Teine Keijinkai Hospital, Sapporo, Japan; [§]Clinical Simulation Center, Hokkaido University Graduate School of Medicine, Sapporo Japan; [‖]Department of Biostatistics, Hokkaido University, Sapporo, Japan; [¶]Department of Surgery, Asahikawa City General Hospital, Asahikawa, Japan; and [#]Department of Surgery, Sapporo Kiyota Hospital, Sapporo, Japan

OBJECTIVE: Item response theory (IRT) was originally developed to make performance assessments more accurate. However, IRT analysis of the intraoperative performance of surgical trainees could help identify the elements that the trainees find difficult during the skill acquisition process. The aim of this study was to identify the essential portions of the skill acquisition process of a surgical procedure using the IRT.

DESIGN: The 24-item assessment checklist was used to evaluate a recorded intra-operative performance of a laparoscopic inguinal hernia repair. The scores were analyzed using IRT to calculate the difficulty and discrimination level of each item.

SETTING: Fifteen institutes.

PARTICIPANTS: Thirty surgical trainees.

RESULTS: A total of 123 assessments were analyzed. The item analysis showed the procedure specific item "traction of peritoneum (difficulty: −0.45; discrimination: 19.37)" and generic items "instrument handling (difficulty: −0.59; discrimination: 3.82)" and "flow of procedure (difficulty: 0.09; discrimination: 3.27)" to be key elements in the skill acquisition process of the procedure.

CONCLUSIONS: Key elements in the skill acquisition process of the procedure were quantitatively identified by applying the IRT analysis. This could lead to the use of IRT in designing and developing a more effective training curriculum. (J Surg Ed 76:1101−1106. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: item response theory analysis, laparoscopic inguinal hernia repair, skill acquisition

ABBREVIATIONS: IRT, Item Response Theory TAPP transabdominal preperitoneal

COMPETENCIES: Patient Care, Practice-Based Learning and Improvement, Systems-Based Practice

INTRODUCTION

In the field of education, most of the examinations we undertake utilize the classical test theory and assess our knowledge level using the total sum of equally weighted questions regardless of their difficulty. This method however fails to reward the examinee who answers difficult questions over the examinee who answers easier questions. Especially for high stake examinations, the educators, have looked at ways to make the assessment process more accurate and effective. Item Response Theory (IRT) is an advanced statistical analysis model often used for this purpose. The Rasch model, is known as a basic IRT model to measure the difficulty parameter of each question. It is then used to give each question its appropriate weight, so that the difficult questions have more marks, thus rewarding the examinees who answer difficult questions over the examinees who answer easier question so that the total marks appropriately reflect the examinees ability.[1]

Correspondence: Inquiries to: Yo Kurashima, MD, PhD, Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Kita 15 Nishi 7, Kita-ku, Sapporo Japan; fax: +81-11-706-7158; e-mail: yo.kurashima@huhp.hokudai.ac.jp

  1931-7204/$30.00
https://doi.org/10.1016/j.jsurg.2019.01.015

However, difficulty alone is not enough to accurately determine the ability of the examinee. Some difficult questions can be easily answered by the examinees with lower ability and vice versa. All the questions have various ability to discriminate between the level of the examinees. Knowledge of the question with higher ability to discriminate between the ability is important to be able to correctly determine the examinees ability. Several new advanced IRT models have been developed over the past few decades considers these factors of difficulty, discrimination ability and also the effect of guessing, in order to overcome the limitations of classic test theory and make the testing process more accurate.[2] They are currently being used for high-stake examinations such as the Test of English as a Foreign Language and the Graduate Management Admission Test.[3]

IRT has also been introduced into the field of medical education, mostly to make the examination process (e.g., medical council qualifying exams) more concise and accurate.[2,4-6] IRT has also been used to recalibrate various assessment scales or optimize the number of questions in existing questionnaires to increase efficiency.[7-11] Although IRT has been used in several fields, applications of IRT in assessing clinical skills are limited. Use of IRT can help optimize the interpretation of a performance assessment by calculating calibrated scores based on the difficulty level and discrimination power of each item.[2] IRT was recently used to calculate the calibrated scores of the Global Operative Assessment of Laparoscopic Skills (GOALS) assessment of laparoscopic procedures such as cholecystectomy, abdominal hernia repair, and colorectal surgery.[12] The original GOALS scores were given by the trained raters assessing the laparoscopic skills in the operation room as the simple sum of the equally weighted scores in 5 domains. IRT calibrated GOALS scores were able to discriminate the difference in ability of the surgeons who had same original score, suggesting IRT calibrated scores had greater assessment precision.

IRT is yet to be used for the analysis of a procedure-specific assessment tool. By applying IRT to a procedure-specific assessment tool, we can obtain data on the difficulty level and discrimination power of each element of the procedure. This can provide insights into the skill acquisition process by identifying the essential elements of the procedure. This could furnish meaningful guidance in the process of designing and developing a procedure-specific training curriculum.

The laparoscopic transabdominal preperitoneal (TAPP) inguinal hernia repair was chosen for this study as it is one of the basic procedures performed by surgical trainees. For the assessment of this procedure, a detailed checklist-type assessment tool has been developed and validity evidence collected using recorded performance.[13] The aim of this study was to identify the essential portions of the skill acquisition process of TAPP procedure by analyzing the TAPP checklist assessment scores using IRT.

METHODS

In this prospective study, we used the video recordings of the intraoperative performance of laparoscopic TAPP inguinal hernia repair by surgical trainees from multiple institutes. Video recordings from the laparoscopic feed without audio were evaluated using the TAPP checklist[13] by blinded raters. All the raters were previously trained in video assessment using the TAPP checklist.

Procedure Specific Assessment Tool: "The TAPP Checklist"

The TAPP checklist divides the TAPP procedure into 18 procedure-specific and 6 generic items (Fig. 1). The performance score was estimated as the sum of the scores of each equally weighted item.[13] Each item was given a score of 1 if the surgeons fulfilled the requirement for that item or 0 if the requirement was not adequately fulfilled. The ability of the surgeon was determined by the total score with higher score indicating better ability. This assessment tool showed good inter-rater reliability among the blinded video raters and the scores correlated with the experience of the surgeons.[13]

Statistical Analysis

The TAPP checklist scores collected were analyzed using the 2-parameter logistic model of IRT. This model estimates difficulty and discrimination scores for each of the 24 items on the TAPP checklist. To apply IRT, the data need to meet the assumption of unidimensionality and local independence.[2] Unidimensionality is the measure of whether the items in the checklist evaluate just 1 item in common. Dimensionality of the data was investigated using polychoric correlations between the items and calculating the eigenvalue of various dimensions of this correlation matrix. A large eigenvalue for first dimension and smaller values for other dimensions indicate unidimensionality of the data. Local independence is the measure of whether each item can independently assess the level of the trainee. It was determined by calculating the residual correlations. Low correlation between the residual of each item indicates that the items are not dependent on each other.

In IRT, the latent ability of the subject is commonly reported on a scale of -4 to 4.[12] The difficulty score of an item is defined as the latent ability score of the subject who has 50% probability of successfully fulfilling the defined requirement, in this case receiving 1 on that item. Subjects require higher latent ability (surgical skill in our case) for a high probability of scoring on items that have a

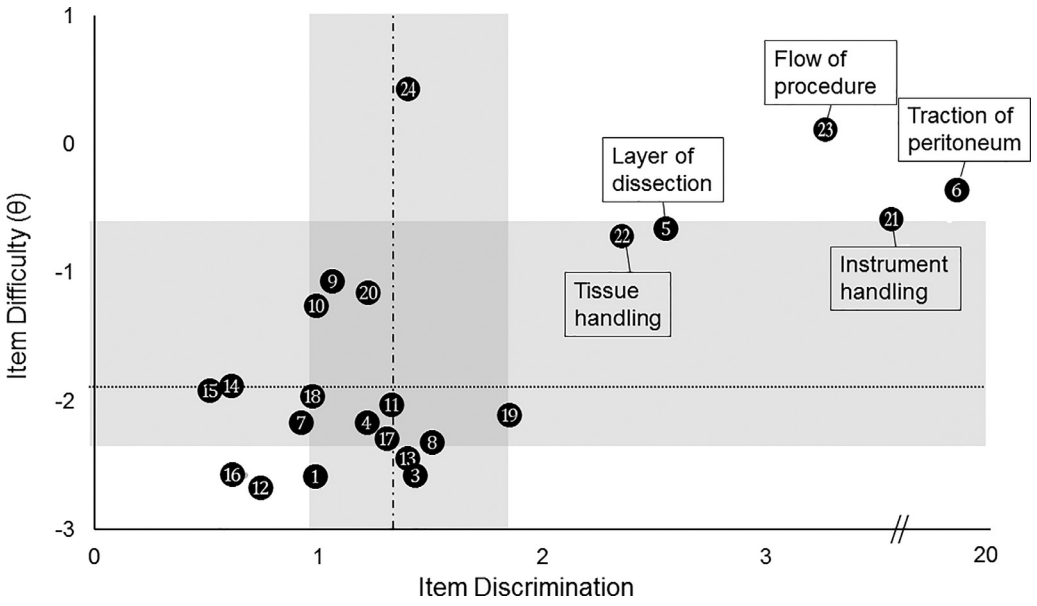

FIGURE 1. Relationship between difficulty and discrimination of each item.
Dotted line represents the median and the shaded area represents the interquartile range.
The numbers indicate the following items of the TAPP Checklist.
Trocar Insertion: 1. view; Incision of Peritoneum: 3. starting point; Creation of Dissection Space; 4. start of dissection, 5. layer of dissection, 6. traction of the peritoneum; Parietalization: 7. safe dissection; Reduction of Hernia Sac: 8. reduction; Extent of Dissection: 9. medial dissection, 10. ventral dissection, 11. lateral dissection, 12. dorsal dissection; Mesh Deployment: 13. size, 14. position, 15. stretching, 16. fixation; Suture of Peritoneum: 17. needle movement, 18. suture final appearance; Overall (generic item): 19. Energy device, 20. Bleeding, 21. Instrument handling, 22. Tissue handling, 23. Flow of procedure, 24. Operation time.

greater difficulty score. Difficulty being a relative item, difficulty scores of each item of the checklist were ranked to determine the most difficult and easiest item in the checklist regardless of their value. Slope parameter is used to calculate the discrimination power of each item. An item with high discrimination value has a greater capability to distinguish between the skill levels. Items with a steep slope require relatively less incremental change in the trainee's ability to increase the probability of scoring on that item. For the purpose of this study, essential items were purposively defined as items with both high difficulty score and discrimination power.

Data are presented as median [interquartile range]. All statistical analyses were performed using JMP version 11 (SAS Institute Inc, Cary, North Carolina) or SAS software version 9.4 (SAS Institute Inc, Cary, North Carolina).

RESULTS

Thirty trainees from 15 institutes underwent the assessment using the TAPP checklist (a median of post graduate year: 4 [3; 8]). A total of 123 TAPP checklist scores were used for the analysis with a median total score of 21 [18; 23], and overall scores ranging from 5 to 24.

The item "trocar position" showed strong negative correlations with 7 other items on the checklist. The eigenvalue of the second dimension was large (4.49), indicating multidimensionality of the data. However, when we reanalyzed the data without the item "trocar position," the eigenvalue of the second dimension decreased. The assumption of unidimensionality was met by eliminating the "trocar position" data. Further IRT analysis was performed without the data of the item "trocar position." The residual correlations of the data were low, with a median of −0.016 [−0.109; 0.070]. This showed the local independence of our data.

The item difficulty and item discrimination parameters for each of the TAPP checklist items are shown in Table 1. The median difficulty score was −1.97 [−2.3; −0.7] and the median discrimination score was 1.31 [0.99; 1.88]. The item analysis showed that "traction of the peritoneum" had the highest discrimination with a score of 19.37, followed by "instrument handling" (3.82) and "flow of the procedure" (3.27). On the difficulty scale, "operation time" was the most difficult item with a score of 0.44, followed by "flow of the procedure" (0.09) and "traction of peritoneum" (−0.45).

The relationship between difficulty and discrimination scores of each item is illustrated in Figure 1. "Traction of peritoneum," "instrument handling," "flow of the procedure," "layer of dissection," and "tissue handling" carried higher scores for both difficulty as well as discrimination power.

TABLE 1. Item Analysis of All the Items of the TAPP Checklist

Item	Difficulty	Discrimination
Procedure-specific items		
Trocar insertion		
1. View	−2.61	1.00
2. Position*	-	-
Incision of peritoneum		
3. Starting point	−2.44	1.40
Creation of dissection space		
4. Start of dissection	−2.20	1.23
5. Layer of dissection†	−0.67	2.58
6. Traction of the peritoneum†	−0.45	19.37
Parietalization		
7. Safe dissection	−2.15	0.93
Reduction of Hernia Sac		
8. Reduction	−2.34	1.51
Extent of dissection		
9. Medial dissection	−1.05	1.06
10. Ventral Dissection	−1.26	0.99
11. Lateral dissection	−2.01	1.33
12. Dorsal dissection	−2.68	0.73
Mesh deployment		
13. Size	−2.54	1.42
14. Position	−1.88	0.64
15. Stretching	−1.89	0.55
16. Fixation	−2.57	0.67
Suture of peritoneum		
17. Needle movement	−2.30	1.31
18. Suture final appearance	−1.97	1.00
Overall (generic item)		
19. Energy device	−2.13	1.88
20. Bleeding	−1.15	1.21
21. Instrument handling†	−0.59	3.82
22. Tissue handling†	−0.70	2.35
23. Flow of procedure†	0.09	3.27
24. Operation time	0.44	1.38

*Omitted for the unidimentiality of the data.
†Essential items.

DISCUSSION

This study identified the key elements of the skill acquisition process of the TAPP procedure by applying IRT. Based on the IRT analysis, procedure-specific items "traction of peritoneum," and "layer of dissection," and generic items "instrument handling," "flow of the procedure," and "tissue handling" were identified as the essential elements of the skill acquisition process of the TAPP procedure for trainees. To the best of our knowledge, this was the first attempt to sample key items from the procedure-specific performance assessment tool by applying IRT. These findings could help educators better understand the portions of the procedure that needs to be focused on when designing and developing the curriculum for the training of the TAPP procedure.

Contrary to our expectations, out of the 5 items that were identified to be essential in this study, 3 were generic items, which means the item is not assessed during one specific part of the procedure, but during the whole procedure. Even "traction of the peritoneum" and "layer of dissection," 2 procedure-specific items that were selected, are generic in nature. This could be because, in this study, the TAPP procedure was mostly performed by laparoscopy inexperienced trainees in a controlled environment. Although they were able to get through the cognitive aspects of the procedures with the help of either various learning tools or their instructors, they found it difficult to overcome their lack of technical expertise. These data support the need for additional training in laparoscopic skills for the trainees outside of the operation room.

One of the main advantages of IRT analysis is that it can calculate both the difficulty level and discrimination power of the items simultaneously. Other methodologies of item analysis can only determine either item difficulty or item discrimination. However, the relationship between the data for item difficulty and item discrimination was significant in determining if the item was truly essential in the skill acquisition process. In our study, taking TAPP procedure as an example, a generic item "operation time," or being able to finish the procedure within the cut-off time, was the most difficult item (0.44). However, it had a relatively low discrimination power (1.38); this means that, while it is a hard item, finishing a procedure within the cut-off time does not necessarily mean that the subject has a higher ability. There were still several subjects with lower performance levels who were able to finish the procedure within the cut-off time. On the other hand, while a procedure-specific item "traction of the peritoneum" is the third most difficult item (−0.45) on the list, it also has high discrimination power (19.37). A trainee scoring on this item was more likely to have a latent ability score of −0.45 or greater, than the ones who did not score on this. "Traction of peritoneum" has a very high discrimination power and is a key item to discriminate the skills being measured.

In the process of data analysis, "trocar position" showed negative correlations with a number of items, affecting the dimensionality of the data. We used only intraoperative video assessment for "trocar position" in this study, and assessment of the item was mainly by assumption of the trocar location based on the angle in which the instrument came into the surgical field. This analysis showed us that, while the trocar position might be an important item for the inexperienced trainees to understand the procedure as a whole and to receive formative feedback, its role as an assessment item is limited.

In the development of the surgical curriculum, significant aspects to be focused on are intuitively determined based on the opinion of an expert panel, or in some cases

from the learners' perspective. Attempts have been made to make this process less subjective using consensus building techniques such as the Delphi method.[14-16] However, these methodologies are based on individual opinions and still have some degree of subjectivity. To obtain truly objective data from the learners, the curriculum should be designed by logically analyzing the performance data. Analyzing the assessments of the trainees using IRT identifies the challenging areas where most trainees struggle during the skill acquisition process. Taking TAPP procedure as an example, with the data, we can say that designing a curriculum that gives more importance to essential items like "traction of peritoneum" is much more likely to help the trainees improve their overall performance than training them how to finish the procedure within the cut-off time. These results, however, do not imply that the trainees need to focus only on "traction of peritoneum" to master the TAPP procedure. The trainees still need to understand the procedure as a whole, as implied by another essential item "flow of the procedure." These results suggest that, while designing an effective curriculum, more detailed information and feedback should be given in these essential areas.

Assessment and feedback have become an integral part of modern surgical education. The past decades have seen the development of various procedure-specific assessment tools.[17,18] Assessment tools when used as part of the training, have played a vital role in developing the technical skills of inexperienced trainees.[19,20] The introduction of IRT can help this process in a number of ways. The IRT data can recalibrate the assessment tool itself to make it more accurate. Mathematical analysis of these assessments can also provide deeper insight into the skill acquisition process of these procedures. This could help surgical educators to develop more practical educational tools or curricula for specific procedures.

This study has several limitations. The trainees performed the surgery in a controlled environment with their instructor ready to takeover or prompt them if the trainee's performance was inadequate. We attempted to minimize this effect by asking the participants to provide us with the information on the portions where the instructors had taken over. However, it was not always possible to evaluate the impact of the intraoperative teaching or prompting on the trainee's performance. No robust evidence exists to support an appropriate sample size for applying the two-parameter logistic model of IRT. Although we determined the important aspects in the skill acquisition process of the TAPP procedure using validated statistical methods, there is no gold standard or consensus on the required sample size for supporting the analysis. We have not studied whether focusing on these aspects will have an actual effect on the training for TAPP procedure for the inexperienced trainees, or whether it will improve the clinical outcome of patients. This was out of the scope of the current study and needs to be confirmed by further studies.

CONCLUSIONS

We analyzed the assessments of the recorded TAPP performance using IRT. We identified procedure specific items "traction of the peritoneum," and "layer of dissection" and generic items "instrument handling," "flow of the procedure," and "tissue handling" as the essential portions of the skill acquisition process of the TAPP procedure. This could help educators deepen their understanding of the skill acquisition process. The findings from this analysis could contribute to the development of an efficient and practical curriculum by facilitating skill acquisition.

ACKNOWLEDGMENTS

We would like to thank all the surgeons who provided the videos of their procedure for this study.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Howard EP. Applying the Rasch Model to test administration. *J Nurs Educ*. 1985;24:340–343.
2. Downing SM. Item response theory: applications of modern test theory in medical education. *Med Educ*. 2003;37:739–745.
3. Lord FM. Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates; 1980.
4. Bhakta B, Tennant A, Horton M, Lawton G, Andrich D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Med Educ*. 2005;5:9.
5. De Champlain AF, Boulais AP, Dallas A. Calibrating the Medical Council of Canada's Qualifying Examination Part I using an integrated item response theory framework: a comparison of models and designs. *J Educ Eval Health Prof*. 2016;13:6.
6. Huang YF, Tsou MY, Chen ET, Chan KH, Chang KY. Item response analysis on an examination in anesthesiology for medical students in Taiwan: a comparison of one- and two-parameter logistic models. *J Chin Med Assoc*. 2013;76:344–349.

7. Benedetti MG, Franchignoni F, Morri M, Franchini N, Natali E, Giordano A. Rasch analysis of the Iowa Level of Assistance Scale in patients with total hip and knee arthroplasty. *Int J Rehabil Res*. 2014;37:118–124.

8. Boeschen Hospers JM, Smits N, Smits C, Stam M, Terwee CB, Kramer SE. Reevaluation of the Amsterdam Inventory for auditory disability and handicap using item response theory. *J Speech Lang Hear Res*. 2016;59:373–383.

9. Crump RT, Liu G, Janjua A, Sutherland JM. Analyzing the 22-item Sino-Nasal Outcome Test using item response theory. *Int Forum allergy Rhinol*. 2016;6:914–920.

10. Engelhard MM, Schmidt KM, Engel CE, Brenton JN, Patek SD, Goldman MD. The e-MSWS-12: improving the multiple sclerosis walking scale using item response theory. *Qual Life Res*. 2016.

11. Fukuhara S, Wakita T, Yamada M, Hiratsuka Y, Green J, Oki K. Development of a short version of the visual function questionnaire using item-response theory. *PloS One*. 2013;8:e73084.

12. Watanabe Y, Madani A, Ito YM, et al. Psychometric properties of the Global Operative Assessment of Laparoscopic Skills (GOALS) using item response theory. *Am J Surg*. 2016.

13. Poudel S, Kurashima Y, Kawarada Y, et al. Development and validation of a checklist for assessing recorded performance of laparoscopic inguinal hernia repair. *Am J Surg*. 2016;212:468–474.

14. Bethlehem MS, Kramp KH, van Det MJ, ten Cate Hoedemaker HO, Veeger NJ, Pierie JP. Development of a standardized training course for laparoscopic procedures using Delphi methodology. *J Surg Educ*. 2014;71:810–816.

15. Miskovic D, Ni M, Wyles SM, et al. 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.

16. May AK, Cuschieri J, Johnson JL, Duane TM, Cherry-Bukowiec JR, Rosengart MR. Determining a core curriculum in surgical infections for fellowship training in acute care surgery using the Delphi technique. *Surg Infect*. 2013;14:547–553.

17. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 2005;138:640–647. discussion 647-649.

18. Zevin B, Bonrath EM, Aggarwal R, et al. Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. *J Am Coll Surg*. 2013;216:955–965. e958; quiz 1029-1031, 1033.

19. Poudel S, Kurashima Y, Kawarada Y, et al. Development of a novel training system for laparoscopic inguinal hernia repair. *Minim Invasive Ther Allied Technol*. 2018. https://doi.org/10.1080/13645706.2018.1504800. In Press.

20. Poudel S, Kurashima Y, Tanaka K, et al. Educational system based on the TAPP checklist improves the performance of novices: a multicenter randomized trial. *Surg Endosc*. 2018;32:2480–2487.