Surg Endosc (2016) 30:512–520
DOI 10.1007/s00464-015-4233-7

Validity evidence for the Fundamentals of Laparoscopic Surgery (FLS) program as an assessment tool: a systematic review

Benjamin Zendejas[1] · Raaj K. Ruparel[1] · David A. Cook[2,3]

Received: 16 December 2014 / Accepted: 8 May 2015 / Published online: 20 June 2015

Abstract
Background The Fundamentals of Laparoscopic Surgery (FLS) program uses five simulation stations (peg transfer, precision cutting, loop ligation, and suturing with extracorporeal and intracorporeal knot tying) to teach and assess laparoscopic surgery skills. We sought to summarize evidence regarding the validity of scores from the FLS assessment.
Methods We systematically searched for studies evaluating the FLS as an assessment tool (last search update February 26, 2013). We classified validity evidence using the currently standard validity framework (content, response process, internal structure, relations with other variables, and consequences).
Results From a pool of 11,628 studies, we identified 23 studies reporting validity evidence for FLS scores. Studies involved residents ($n = 19$), practicing physicians ($n = 17$), and medical students ($n = 8$), in specialties of general ($n = 17$), gynecologic ($n = 4$), urologic ($n = 1$), and veterinary ($n = 1$) surgery. Evidence was most common in the form of relations with other variables ($n = 22$, most often expert–novice differences). Only three studies reported internal structure evidence (inter-rater or inter-station reliability), two studies reported content evidence (i.e., derivation of assessment elements), and three studies reported consequences evidence (definition of pass/fail thresholds). Evidence nearly always supported the validity of FLS total scores. However, the loop ligation task lacks discriminatory ability.
Conclusion Validity evidence confirms expected relations with other variables and acceptable inter-rater reliability, but other validity evidence is sparse. Given the high-stakes use of this assessment (required for board eligibility), we suggest that more validity evidence is required, especially to support its content (selection of tasks and scoring rubric) and the consequences (favorable and unfavorable impact) of assessment.

Keywords Validation · Fundamentals of Laparoscopic Surgery · Simulation · Assessment

The Fundamentals of Laparoscopic Surgery (FLS) program is an educational system designed to teach and assess the fundamental knowledge and technical skills required in basic laparoscopic surgery. FLS evolved from the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) [1] and presently includes web-based didactics, hands-on skills training, and an assessment tool [2]. With endorsement from the Society of American Gastrointestinal and Endoscopic Surgeons and the American College of Surgeons, both the training and assessment portions of the FLS are now required by the American Board of Surgery as prerequisites to board eligibility in general surgery, and as such the FLS program impacts virtually every general surgery resident in the USA.

This article was presented at the 2015 Annual Meeting of the Society of American Gastrointestinal and Endoscopic Surgeons in Nashville, TN, on Friday, April 17, 2015 (Abstract ID 62485, Session Number SS18). Presented at the SAGES 2014 Annual Meeting, April 2-5, 2014, Salt Lake City, Utah.

✉ Benjamin Zendejas
zendejas.benjamin@mayo.edu

[1] Department of Surgery, Mayo Clinic College of Medicine, Mayo 12-W, 200 First Street SW, Rochester, MN 55905, USA

[2] Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA

[3] Multidisciplinary Simulation Center, Mayo Clinic College of Medicine, Rochester, MN, USA

Several studies confirm the training effectiveness of the FLS program [3]. However, the assessment component of the FLS has received relatively less research attention. Despite this, given its role in the credentialing process, it has become a de facto high-stakes test [4]. The implications of such high-stakes assessment, and the corresponding decisions, on the lives of individual trainees suggest the need for strong evidence to support the validity of scores and the defensibility of their proposed interpretations and uses [5, 6]. Unfortunately, reviews describing the validity evidence surrounding the use of FLS as an assessment instrument are limited by their age, nonsystematic inclusion of studies, or lack of a comprehensive validity framework [3, 7]. More importantly, recommendations to address the weakest links in the validity evidence chain are missing.

We sought to systematically review the current state of validity evidence surrounding the use of FLS as an assessment instrument. We placed special emphasis on interpreting prior validation efforts in light of the validity framework first proposed by Messick [8], advocated by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) in the 1999 *Standards for Educational and Psychological Testing* [9], and reaffirmed in the 2014 *Standards* [10]. This framework regards validity as a hypothesis supported by evidence derived from five sources, namely content, response process, internal structure, relations with other variables, and consequences [8, 9]. We further sought to highlight the gaps in the validity evidence chain so that further research may seek to complete the validity argument to support or refute the proposed score interpretations and uses of the FLS assessment. Finally, we sought to identify barriers to and variations in implementation of the FLS assessment in practice.

Methods

This review was planned, conducted, and reported in adherence to PRISMA standards of quality for reporting systematic reviews [11]. We previously reported on studies that have used technology-enhanced simulation to assess health professionals [12]. The present topic-focused review emerged from that broad overview, but we collected new data in order to delve deeply into the included studies and thereby gain new insights. We describe briefly the methods reported previously and highlight those unique to the present report. Because MISTELS was the backbone upon which the tasks and assessment portions of the FLS evolved from, we purposefully review MISTELS-related validity evidence as it directly impacts the FLS validation argument. In the interest of simplicity, we will refer to both the FLS and its predecessor, the MISTELS, as "FLS."

We sought to answer the questions

- How well does published evidence support or refute the validity of scores for the FLS assessment?
- What barriers have been reported during the implementation of the FLS assessment?

Evaluating the validity of education assessments

We coded the prevalence of each of the five evidence sources noted above (see Table 1 for definitions, and see our prior work for detailed elaborations and examples) [13]. We evaluated validity evidence separately for the knowledge and skills components of the FLS assessment.

Study eligibility

We included all studies that evaluated the FLS assessment of laparoscopic skills in health professional learners, at any stage in training or practice, using technology-enhanced simulation [14]. To be eligible, studies had to provide new data or evidence that either directly or indirectly supported or refuted the interpretation of FLS scores (as compared with summarizing previously published data, or describing FLS use without reporting validity evidence).

Study identification

Our search strategy has been previously published in full [12, 14, 15]. In brief, we searched multiple databases, including MEDLINE, ERIC, and Scopus, for relevant articles using a search strategy developed by an experienced research librarian. Examples of search terms included simulat*, assess*, evaluat*, valid*,and reliab*. We used no beginning date cutoff, and the last date of search was February 26, 2013.

Study selection

Working in duplicate, reviewers screened all candidate studies for inclusion. We first reviewed each title and abstract; then, if needed, we reviewed the full text of studies judged eligible or uncertain. We resolved conflicts by consensus. Inter-rater agreement for study inclusion was substantial [intraclass correlation coefficient (ICC) = 0.72].

Data extraction and synthesis

We developed a data abstraction form through iterative testing and revision. We abstracted data independently and in duplicate for validity evidence sources as outlined above, resolving conflicts by consensus. In coding for

Table 1 Operational definition of validity elements

Evidence element	Definition
Content	Steps taken to ensure that test content reflects the construct it is intended to measure
Internal structure—reliability	Reproducibility of scores across
Inter-rater	Different raters
Inter-station	different stations or tasks
Test–retest	different versions of the test
Relations with other variables	
Learner characteristic	Association with training level (expert/novice) or status (trained/untrained)
Training responsiveness	Change in scores following training interventions
Separate measure	Association with a separate measure, with a hypothesized relation with test scores
Response process	Analysis of raters' thoughts/actions while scoring; test security, quality control
Consequences	Impact, beneficial or harmful, of the assessment itself

See Cook and Beckman [5], Cook et al. [13], and the AERA, APA, NCME standards for more information [8, 10]

validity evidence, our inter-rater agreement ICCs (as reported previously) [12] ranged 0.67–0.91 except for response process (ICC = 0.34, raw agreement 95 %) and consequences (ICC = 0.56). As per Landis and Koch [16], inter-rater agreement values 0.21–0.4 are considered "fair," 0.41–0.6 "moderate," 0.61–0.8 "substantial," and >0.8 "almost perfect." In addition to reporting means and counts, we conducted a critical synthesis of findings to identify advantages, barriers, and modifications to implementing the FLS assessment in practice. We graded study quality with the Medical Education Research Study Quality Instrument (MERSQI) [17].

Results

Trial flow and study characteristics

From a pool of 11,628 potentially relevant articles, we included 23 studies (Fig. 1), enrolling 1280 participants (median 40 participants per study, range 12–215). The first description was in 1998 [1], but the majority of reports ($n = 13$, 57 %) were published in or after 2007. With the exception of one study that involved veterinary medicine students [18], studies involved physicians at some stage of training including postgraduate physician trainees (i.e., residents; $n = 19$ studies, 83 %), practicing physicians ($n = 17$, 74 %), and medical students ($n = 8$, 35 %). Nineteen (83 %) studies included participants from more than one training stage. Eight (35 %) studies were multi-institutional. Participant specialties included general surgery ($n = 17$, 74 %), obstetrics/gynecology ($n = 4$, 17 %), and urology and veterinary surgery ($n = 1$, 4 % each). Median MERSQI (range) scores were 9.5 (6.5–11), out of a maximum possible score of 18.

Validity evidence for the FLS skills assessment

No studies interpreted their validation argument using the currently standard validity framework [9, 10]. Rather, studies used either the classical validity framework (content, criterion [concurrent and predictive], and construct validity; $n = 8$, 35 %) or a more limited framework, such as construct validity alone ($n = 10$, 43 %). Five studies reported no validity framework whatsoever. Most studies ($n = 15$, 65 %) reported only one validity evidence source, six reported two sources, and two reported three.

Content evidence

Though most studies provided some description of the FLS tasks and their respective scoring metrics, only two studies have reported actual evidence evaluating the match between FLS tasks and scores and the target construct (i.e., content evidence). Fried et al. described the initial selection of FLS tasks as being determined by consensus among five expert laparoscopic surgeons, although the method of achieving consensus was not defined. Fraser et al. [3] later proposed a revised scoring process in which scores are standardized against the performance of surgery chief residents (postgraduate year 5) and normalized so that all task scores ranged from 0 to 100 (as opposed to having different ranges for each task) [19].

Internal structure evidence

Only three studies have evaluated internal structure of FLS scores, namely inter-station reliability ($N = 2$), test–retest reliability ($N = 2$), and inter-rater reliability ($N = 1$). All these reliability estimates supported the validity of FLS scores (ICC $\geq$ 0.77) [20–22].

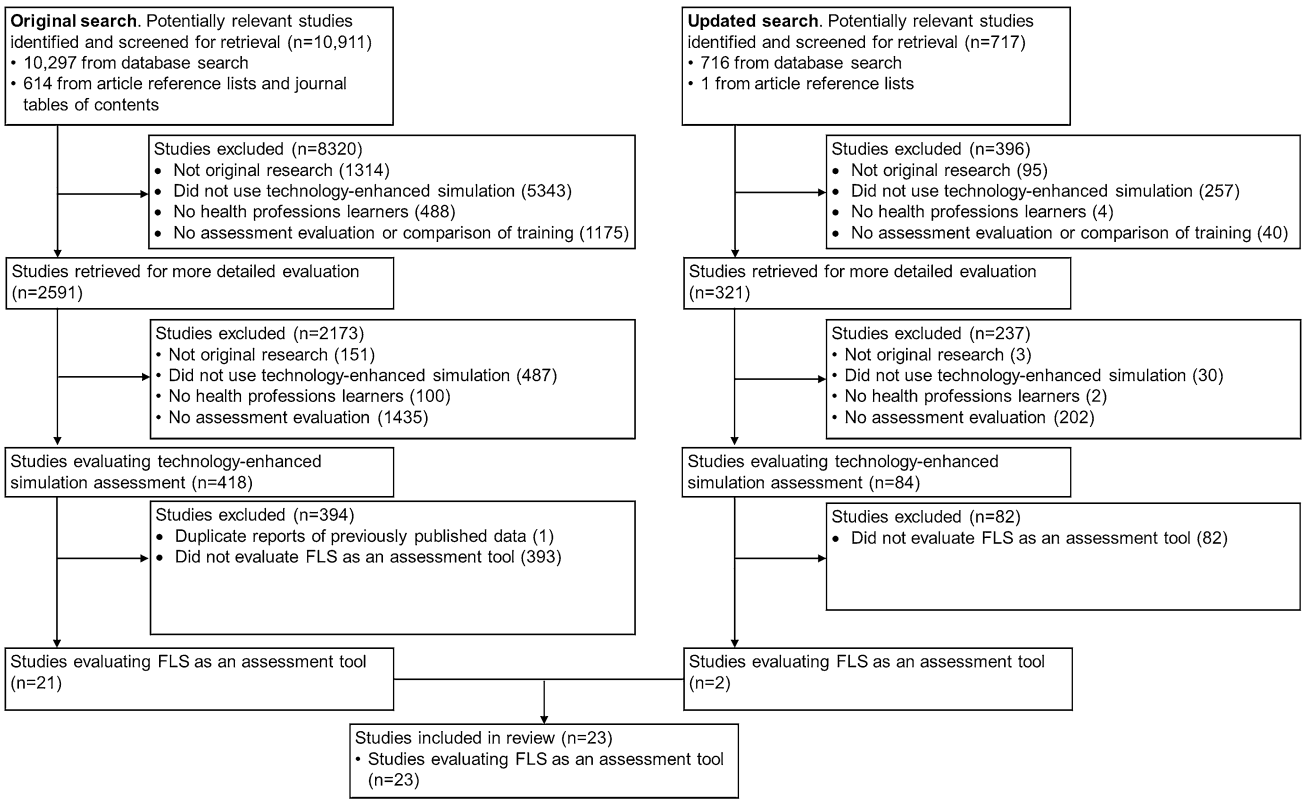

Fig. 1 Flow of included studies

Relations with other variables evidence

An important source of validity evidence comes from evaluations of relations with other variables such as training status or another assessment score. Seventeen (74 %) studies showed that total FLS skills scores discriminate between levels of training [usually different postgraduate years (see Table 2)], which offers weak but supportive evidence of validity [23].

FLS total skills scores have shown sensitivity to change, with significant score improvements after FLS and nonFLS simulation-based training in laparoscopic surgery [1, 3, 18, 21, 24–26]. Furthermore, training on a basic task (peg transfer) has shown to independently improve scores of a related, though more advanced, task (intracorporeal suturing) [3]. Skill scores have also shown expected performance decrements in response to situations associated with greater mental workload caused by dual-task working conditions or transfer tasks [26, 27].

FLS total skill scores have been shown to favorably correlate with other assessment metrics, such as in-training assessments by supervising physicians (rotation grades, $r = 0.51$) [28], automated scoring on another simulator (detailed below), operative performance assessments (GOALS) [29] during laparoscopic cholecystectomies ($r = 0.77$–0.81), [3, 30] prior laparoscopic experience (i.e., case volume, $r = 0.55$) [31, 32], and with self-reported laparoscopic skills (confidence, $r = 0.59$) [24]. Considering specifically the studies of automated scoring, these have shown statistically significant positive correlations between the FLS skills scores and the augmented-reality ProMIS simulator computer-generated metrics of path length and instrument smoothness ($r = 0.5$–0.86 and $r = 0.94$–0.99, respectively) [21], the LTS 2000-ISM60 and VBlast simulator scores based on speed and precision ($r = 0.79$, and $r = 0.36$, respectively) [22, 33], motion sensor metrics of number of movements and path distance (ICSAD, $r = 0.76$ and $r = 0.81$, respectively) [34], and grasp force and work-sphere volume (SIMIS, $r = 0.51$ and $r = 0.58$, respectively) [35]. Similar favorable correlations have been shown when FLS tasks and their respective assessments have been adapted to nonstandard applications including animal (porcine) models [24], robotics [36], and pediatric surgery [37].

Response process evidence

The only response process evidence for FLS scores comes from two studies that standardized the responses of raters through training, rigorous assessment of rating accuracy (blinded video review), and rater retraining as needed [3, 32].

Table 2 Validity evidence for FLS skills assessment

Author (year)	Training level (specialty)	*N*	MERSQI	Validity evidence—FLS skills						
				Content	Internal structure	Relations with other variables			Response process	Consequences
						Experience level	Training responsiveness	Concurrent/predictive assessments		
DeRossis et al. [1]	PGY, MD (GS)	46	8.5			▲ ○	▲			
Fried and Derossis [13]	PGY (GS)	12	10				▲ ○	▲ ○		
Fraser et al. [19]	MS, PGY, MD (GS)	165	8.5	▲		▲ ○				▲
Feldman et al. [28]	PGY (GS)	50	9.5					▲ ○		
Fried et al. [3]	PGY, MD (GS)	215	10.5	▲		▲	▲	▲	▲	
Dauster et al. [38]	PGY, MD (Urology)	17	8			▲ ○				
Avgerinos [48]	MS, PGY (GS)	32	10			▲				
Fichera [49]	PGY, MD(OB/GYN)	40	9.5			▲				
Vassiliou et al. [20]	MS, PGY, MD (GS)	12	9.5		▲					
Swanstrom et al. [32]	PGY, MD (GS)	70	10			▲		▲	▲	
Stefanidis et al. [27]	PGY, MD (GS)	27	9.5			▲	▲			
Ritter et al. [21]	MS, PGY, MD (GS)	60	8.5		▲	▲	▲	▲		
McCluney et al. [30]	PGY, MD (GS)	40	9.5					▲		▲
Kolkman et al. [25]	MS, MD (OB/GYN)	23	10			▲ ○	▲			
Xeroulis et al. [34]	PGY, MD (GS)	26	9			▲		▲		
Sansregret et al. [22]	MS, PGY, MD (OB/GYN)	111	10		▲	▲		▲		
Sankaranarayanan et al. [33]	PGY, MD (GS)	50	9.5			▲ ○		▲		
Zheng et l. [31]	PGY, MD (OB/GYN)	41	9.5			▲				
Yurko et al. [26]	MS (GS)	28	7.5				▲	▲		
Fransson and Ragle [18]	VS (Veterinary)	33	11			▲	▲			
Azzie et al. [37]	MD (GS)	50	10			▲		▲ ○		
Jayaraman et al. [35]	PGY, MD (GS)	15	9					▲		▲
Stefanidis et al. [36]	PGY, MD (GS)	117	9.5			▲		▲ ○		
Totals		1280	9.5[a]	▲2	▲3	▲17, ○ 5	▲8, ○ 1	▲13, ○ 4	▲2	▲3

FLS skills = Fundamentals of Laparoscopic Surgery skills, (▲) = positive evidence favoring FLS skill scores, (○) = negative evidence against FLS skill scores, PGY = resident or postgraduate trainee, MS = medical student, MD = practicing physician or fellow, VS = veterinary student, OB/GYN = obstetrics and gynecology, GS = general surgery, MERSQI = Medical Education Research Study Quality Instrument (maximum possible score = 18) [17]. [a] Median. N = number of participants being assessed

Consequences evidence

No studies have directly evaluated the intended or unintended consequences of the FLS assessment itself (i.e., does the assessment—perhaps associated with remediation for low performers—have beneficial or harmful impacts on trainees or patients?). An indirect form of consequences evidence can derive from studies that rigorously establish a pass–fail standard; we found three such studies, all of which used a receiver operating curve (ROC) to establish the passing score [19, 29, 30].

Task-specific evidence

For the most part, task-specific scores have shown discriminatory and correlational characteristics similar to those described above for total FLS scores, with two notable exceptions. First, the scores from the MISTELS clip application and mesh placement tasks showed non-significant correlations with level of training [1] and with skill scores in a porcine model; [24] these tasks were not included in the FLS program. Second, several studies have found that the scores from the MISTELS/FLS loop ligation task fail to discriminate performance across levels of trainee experience [1, 24, 25, 33, 38].

Validity evidence for the FLS knowledge assessment

Far less evidence has been reported for the FLS knowledge assessment. Content evidence for the knowledge-based assessment portion of the FLS is provided exclusively by Swanstrom et al. [32], who report questions being developed through expert consensus and test blueprint. This study also provides the only internal-structure-type evidence for the knowledge portion of the FLS assessment, with an ICC of 0.81 (for an average of 68 questions per test), and further reports relations with other variables in the form of statistically significant positive correlations ($r = 0.56$–0.76) between FLS knowledge scores and FLS skills scores, prior laparoscopic experience, levels of training, and confidence [32]. Conversely, another study found that FLS knowledge scores did not significantly correlate with levels of training, confidence, or prior laparoscopic experience for gynecologic surgeons [31]. No evidence of response process or consequences has been reported for the FLS knowledge component.

Barriers to implementation and adaptation

Subjectivity of interpretation, lack of immediate scoring and feedback, and cost have been listed as disadvantages of the FLS assessment program [22, 32]. However, we found no data to empirically indicate the frequency or severity of these barriers.

Azzie et al. described some of the difficulties encountered as they adapted the FLS program to the training and assessment of pediatric surgery-specific skills. Most notably, they found that using the same FLS scoring rubric to score conceptually similar, but physically smaller tasks was not ideal as it led to lack of discriminatory ability for certain tasks. They suggested that such "smaller" tasks were of either increased difficulty because of less space to maneuver or completed more quickly because of the smaller surface area involved [37].

Discussion

The FLS program is one of the most successful and widespread simulation programs in the history of technology-enhanced simulation. While its training effectiveness has been well demonstrated, the evolution of its assessment methodology and the corresponding score validation efforts leave room for improvement. This review demonstrates that evidence supporting the interpretation of the FLS scores is incomplete, arising predominantly from studies evaluating relations with other variables such as known-group comparisons, responsiveness to training, and correlation with other measures. Evidence supporting the content, internal structure, response process, and consequences of scores is sparse. With the insight gained from this review, we propose the following specific recommendations to address such gaps.

First, further content and internal structure (particularly inter-rater and inter-task reliability) evidence should be collected. The paucity of evidence to support the content (e.g., task selection and scoring) of the FLS assessment is concerning. Content evidence serves as the foundation upon which all other sources of evidence rest, and with a weak foundation it becomes harder to support the remaining sources of evidence. While it is obviously too late to justify the initial selection of the specific tasks or the scoring metric and expert surgeons may have a hard time disregarding their knowledge of the FLS in order to objectively define ideal tasks de novo, evidence could nonetheless be gathered to explore the match (or mismatch) between FLS tasks and the essential skills required by beginning laparoscopists. For example, important skills may be poorly represented in the current FLS tasks, or current tasks may need modification to ideally reflect best practices. The scoring rubric also warrants rigorous evaluation. Evidence of internal structure is generally supportive of FLS scores, but has been reported in only a few studies, and these are limited by small sample sizes and rely on concomitant evaluation of known-group

comparisons which typically overestimate reliability indices [23]. Further evaluation of the reliability of FLS scores, perhaps using generalizability theory [39, 40], would provide additional internal structure evidence to strengthen the validity argument.

Second, we must stop relying on expert–novice comparisons to justify FLS validation. Though the *lack* of such discriminatory ability would raise concerns (as has been noted for some FLS subtasks) [1, 24, 25, 33, 38], *confirmation* of expert–novice differences does little to advance the validity argument [23]. The main problem stems from the multitude of possible reasons that may account for such observed differences. Association does not imply causation, and we simply cannot assume that the observed differences are a result of only the expert's degree of laparoscopic skill. In addition, such expert–novice comparisons suffer from spectrum bias [41], in which groups artificially created to reflect the extremes of performance show spuriously inflated discriminatory power. In typical practice learning groups are far less heterogeneous and discrimination is more difficult.

Third, it is time to seriously reconsider the value of the loop ligation task, which has poor discriminatory ability and yet is the most expensive FLS task currently used. Tasks with similarly poor discrimination in the MISTELS program were not included in the current FLS. We suggest that barring new evidence supporting the validity of loop ligation scores (such as content evidence suggesting that it is a critical skill independent of other FLS tasks, or evidence confirming meaningful relations with other variables), it may be time to drop this task from the FLS assessment.

After we have addressed the above-mentioned recommendations, priority should be given to the collection of evidence of consequences. So far, such evidence has only been provided in the form of pass–fail thresholds. Though it is important to be able to distinguish those who pass or fail a test, more direct and meaningful evidence of consequences will come from studies exploring the anticipated and unanticipated effects of FLS testing. Such evidence will be particularly important as FLS scores become an integral part of the residency milestone project [42, 43] with a corresponding requirement for remediation for those who do not pass. Our prior work lists examples of consequences evidence that could be collected [13].

We do not imply that FLS scores are either valid or invalid, nor do we recommend the abandonment of this tool. We view ourselves as detectives collecting and presenting evidence rather than the judge and jury. Different stakeholders (educators, administrators, etc.) will interpret the evidence differently and may arrive at different conclusions. However, the issues we have identified will require attention regardless of the validity framework used, and it remains to be seen how these issues will be addressed. Fortunately, validity is not static. An assessment instrument (in this case the FLS) does not receive a permanent stamp of valid or invalid. Rather, validity evidence can continue to accrue to support or refute the proposed interpretations of assessment scores.

We anticipate that our work will help guide the validation efforts of novel assessment instruments which are following the FLS footsteps, such as the Fundamentals of Endoscopic Surgery [44], the Fundamentals of Use of Surgical Energy [45], and the Fundamentals of Robotic Surgery (FRS) [46]. Carefully planning a validity argument and then strategically collecting evidence to test the most important or questionable assumptions in that argument will allow educators and other stakeholders to judge the defensibility of decisions based on assessment scores. Validation efforts that are deliberate, comprehensive, and well balanced will advance the science of assessment in surgical education.

Limitations

This review has limitations. We had suboptimal inter-rater agreement for response process evidence. This could be due to ambiguous reporting (for example, authors rarely identified evidence of consequences or response process as such) or imprecise definitions, which highlights the need for greater clarity in the definitions for these elements. Nonetheless, we reached consensus on all reported data. Also, for studies that reported more than one assessment instrument we abstracted information only for that pertaining to the FLS assessment. We cannot exclude the possibility that studies showing nonsignificant or unfavorable evidence remain unpublished (i.e., publication bias), particularly for studies exploring associations with clinical outcomes.

Although the framework we used for interpreting evidence was not used in any of the original studies, the framework itself is not new [8–10]. More importantly, most of the concepts [content evidence, relations with other variables (construct and criterion validity), and internal structure (e.g., reliability)] have been around for over 50 years. We have simply applied a new, more comprehensive lens that facilitates integrating evidence from disparate sources and identifying evidentiary gaps. Whether we use an old framework or a modern one, the validity evidence for the FLS assessment clearly remains incomplete. Moreover, if the validity evidence for any assessment's scores is insufficient, then the ongoing use of that instrument's scores for high-stakes decisions must be carefully considered. However, absence of evidence is not the same as evidence of absence; collection of additional evidence may yet support the use of the FLS assessment as a high-stakes test.

Comparisons with previous reviews

The findings of the present review are congruent with those of previous reviews of simulation-based assessments [12], clinical skills assessments [47], and the FLS assessment program [3, 7]. For example, our recent systematic review of 417 studies of simulation-based assessments found a similar pattern of evidence, with known-group or expert–novice comparisons being the most common single source of validity evidence (73 % of studies) and the only source of validity evidence in a third of the studies [12]. However, contrary to a review of observational skills assessment instruments in which most of the validity evidence came from evaluations of novice or trainee level learners [47], our findings suggest that a substantial proportion of studies evaluating the FLS assessment included learners from all stages of training, including practicing physicians. Median MERSQI scores [17] for studies evaluating the FLS program were substantially lower than those of other assessment instruments [12], suggesting room for improvement in the design, conduct, and reporting of studies evaluating the FLS. We are unaware of studies that have interpreted FLS validation efforts using the currently accepted validation framework [10], for which we believe that our work represents a unique contribution to the field.

Conclusion

The FLS assessment program has accrued substantial validity evidence supporting relations with other variables. However, strong evidence from one source cannot compensate for lack of vital evidence from other sources. Our review highlights that important evidence gaps exist and as such the validity argument for FLS scores remains incomplete. While it may seem unfair to use a broad and relatively new validity framework to interpret studies that used an older, narrower framework, our systematic approach allowed us to identify gaps that would otherwise be missed. Our intent is not to discredit or disparage prior work, but to reveal areas requiring future attention in efforts to validate the use of FLS scores for high-stakes assessment.

Disclosure Drs. Zendejas, Ruparel, and Cook have no conflicts of interest or financial ties to disclose.

References

1. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL (1998) Development of a model for training and evaluation of laparoscopic skills. Am J Surg 175:482–487
2. Fundamentals of laparoscopic surgery. http://www.flsprogram.org/
3. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, Andrew CG (2004) Proving the value of simulation in laparoscopic surgery. Ann Surg 240:518–525
4. ABS to Require ACLS, ATLS and FLS for General Surgery Certification. http://www.absurgery.org/default.jsp?news_newreqs
5. Cook DA, Beckman TJ (2006) Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 119(166):e7–e16. doi:10.1016/j.amjmed.2005.10.036
6. Downing SM (2003) Validity: on meaningful interpretation of assessment data. Med Educ 37:830–837
7. Vassiliou MC, Dunkin BJ, Marks JM, Fried GM (2010) FLS and FES: comprehensive models of training and assessment. Surg Clin N Am 90:535–558. doi:10.1016/j.suc.2010.02.012
8. Messick S (1989) Validity. In: Linn RL (ed) Educational Measurement, 3rd edn. American Council on Education and Macmillan, New York, pp 13–103
9. American Educational Research Association—American Psychological Association & National Council on Measurement in Education (1999) Standards for educational and psychological testing. Washington, DC
10. American Educational Research Association—American Psychological Association & National Council on Measurement in Education (2014) Standards for educational and psychological testing. Washington, DC
11. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151(264–9):W64
12. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R (2013) Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med 88:872–883. doi:10.1097/ACM.0b013e31828ffdcf
13. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R (2014) What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract 19:233–250. doi:10.1007/s10459-013-9458-4
14. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ (2011) Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 306:978–988
15. Brydges R, Hatala R, Zendejas B, Erwin P, Cook D (2015) Linking Simulation-Based Educational Assessments and Patient-Related Outcomes: A Systematic Review and Meta-Analysis. Acad Med. 90(2):246–256. doi:10.1097/ACM.0000000000000549
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
17. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM (2007) Association between funding and quality of published medical education research. JAMA 298:1002–1009. doi:10.1001/jama.298.9.1002
18. Fransson BA, Ragle CA (2010) Assessment of laparoscopic skills before and after simulation training with a canine abdominal model. J Am Vet Med Assoc 236:1079–1084. doi:10.2460/javma.236.10.1079
19. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM (2003) Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc 17:964–967. doi:10.1007/s00464-002-8828-4
20. Vassiliou MC, Ghitulescu GA, Feldman LS, Stanbridge D, Leffondré K, Sigman HH, Fried GM (2006) The MISTELS program to measure technical skill in laparoscopic surgery: evidence for reliability. Surg Endosc 20:744–747. doi:10.1007/s00464-005-3008-y
21. Ritter EM, Kindelan TW, Michael C, Pimentel EA, Bowyer MW (2007) Concurrent validity of augmented reality metrics applied

to the fundamentals of laparoscopic surgery (FLS). Surg Endosc 21:1441–1445. doi:10.1007/s00464-007-9261-5
22. Sansregret A, Fried GM, Hasson H, Klassen D, Lagacé M, Gagnon R, Pooler S, Charlin B (2009) Choosing the right physical laparoscopic simulator? Comparison of LTS2000-ISM60 with MISTELS: validation, correlation, and user satisfaction. Am J Surg 197:258–265. doi:10.1016/j.amjsurg.2008.02.008
23. Cook DA (2014) Much ado about differences: Why expert-novice comparisons add little to the validity argument. Adv Health Sci Educ Theory Pract. [Epub ahead of print]
24. Fried GM, Derossis AM, Bothwell J, Sigman HH (1999) Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. Surg Endosc 13: 1077–1081 **discussion 1082**
25. Kolkman W, Put MAJ, Wolterbeek R, Trimbos JBMZ, Jansen FW (2007) Laparoscopic skills simulator: construct validity and establishment of performance standards for residency training. Gynecol Surg 5:109–114. doi:10.1007/s10397-007-0345-y
26. Yurko YY, Scerbo MW, Prabhu AS, Acker CE, Stefanidis D (2010) Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. Simul Healthc 5:267–271. doi:10.1097/SIH.0b013e3181e3f329
27. Stefanidis D, Scerbo MW, Korndorffer JR, Scott DJ (2007) Redefining simulator proficiency using automaticity theory. Am J Surg 193:502–506. doi:10.1016/j.amjsurg.2006.11.010
28. Feldman LS, Hagarty SE, Ghitulescu G, Stanbridge D, Fried GM (2004) Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. J Am Coll Surg 198:105–110. doi:10.1016/j.jamcollsurg.2003.08.020
29. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113. doi:10.1016/j.amjsurg.2005.04.004
30. McCluney AL, Vassiliou MC, Kaneva PA, Cao J, Stanbridge DD, Feldman LS, Fried GM (2007) FLS simulator performance predicts intraoperative laparoscopic skill. Surg Endosc 21:1991–1995. doi:10.1007/s00464-007-9451-1
31. Zheng B, Hur H-C, Johnson S, Swanström LL (2010) Validity of using fundamentals of laparoscopic surgery (FLS) program to assess laparoscopic competence for gynecologists. Surg Endosc 24:152–160. doi:10.1007/s00464-009-0539-7
32. Swanstrom LL, Fried GM, Hoffman KI, Soper NJ (2006) Beta test results of a new system assessing competence in laparoscopic surgery. J Am Coll Surg 202:62–69. doi:10.1016/j.jamcollsurg.2005.09.024
33. Sankaranarayanan G, Lin H, Arikatla VS, Mulcare M, Zhang L, Derevianko A, Lim R, Fobert D, Cao C, Schwaitzberg SD, Jones DB, De S (2010) Preliminary face and construct validation study of a virtual basic laparoscopic skill trainer. J Laparoendosc Adv Surg Tech A 20:153–157. doi:10.1089/lap.2009.0030
34. Xeroulis G, Dubrowski A, Leslie K (2009) Simulation in laparoscopic surgery: a concurrent validity study for FLS. Surg Endosc 23:161–165. doi:10.1007/s00464-008-0120-9
35. Jayaraman S, Trejos AL, Naish MD, Lyle A, Patel RV, Schlachta CM (2011) Toward construct validity for a novel sensorized instrument-based minimally invasive surgery simulation system. Surg Endosc 25:1439–1445. doi:10.1007/s00464-010-1411-5
36. Stefanidis D, Hope WW, Scott DJ (2011) Robotic suturing on the FLS model possesses construct validity, is less physically demanding, and is favored by more surgeons compared with laparoscopy. Surg Endosc 25:2141–2146. doi:10.1007/s00464-010-1512-1
37. Azzie G, Gerstle JT, Nasr A, Lasko D, Green J, Henao O, Farcas M, Okrainec A (2011) Development and validation of a pediatric laparoscopic surgery simulator. J Pediatr Surg 46:897–903. doi:10.1016/j.jpedsurg.2011.02.026
38. Dauster B, Steinberg AP, Vassiliou MC, Bergman S, Stanbridge DD, Feldman LS, Fried GM (2005) Validity of the MISTELS simulator for laparoscopy training in urology. J Endourol 19:541–545. doi:10.1089/end.2005.19.541
39. Cook DA, Beckman TJ, Mandrekar JN, Pankratz VS (2010) Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. Adv Health Sci Educ 15:633–645. doi:10.1007/s10459-010-9224-9
40. Bloch R, Norman G (2012) Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. Med Teach 34:960–992. doi:10.3109/0142159X.2012.703791
41. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM (1999) Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282: 1061–1066
42. Nasca TJ, Philibert I, Brigham T, Flynn TC (2012) The next GME accreditation system—rationale and benefits. N Engl J Med 366:1051–1056. doi:10.1056/NEJMsr1200117
43. Cogbill TH, Ashley SW, Borman KR, Buyske J, Cofer JB, Deladisma AM (2014) The General Surgery Milestone Project. http://www.acgme.org/acgmeweb/tabid/150/ProgramandInstitutionalAccreditation/SurgicalSpecialties/Surgery.aspx. Accessed 27 May 2015
44. Fundamentals of endoscopic surgery. http://www.fesprogram.org/. Accessed 5 May 2015
45. Fundamental Use of Surgical Energy. http://www.fuseprogram.org/. Accessed 5 May 2015
46. Fundamentals of robotic surgery. http://frsurgery.org/. Accessed 5 May 2015
47. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB (2011) Observational tools for assessment of procedural skills: a systematic review. Am J Surg 202(469–480):e6. doi:10.1016/j.amjsurg.2010.10.020
48. Avgerinos DV, Goodell KH, Waxberg S, Cao CGL, Schwaitzberg SD (2005) Comparison of the sensitivity of physical and virtual laparoscopic surgical training simulators to the users level of experience. Surg Endosc 19:1211–1215. doi:10.1007/s00464-004-8256-8
49. Fichera A, Prachand V, Kives S, Levine R, Hasson H (2005) Physical Reality Simulation for Training of Laparoscopists in the 21st Century. A Multispecialty, Multi-institutional Study. JSLS 9:125–129