Validation of measures from a thoracoscopic esophageal atresia/tracheoesophageal fistula repair simulator

Katherine A. Barsness a,b,⁎, Deborah M. Rooney c, Lauren M. Davis b, Anthony C. Chin a

a Division of Pediatric Surgery, Ann and Robert H Lurie Children’s Hospital of Chicago, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
b Center for Simulation Technology and Immersive Learning, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
c Department of Medical Education, University of Michigan Medical School, Ann Arbor, MI, USA

A R T I C L E   I N F O

Article history:
Received 11 September 2013
Accepted 30 September 2013

Key words:
Thoracoscopic
Tracheoesophageal Fistula
Simulation
Validation

A B S T R A C T

Purpose: A validated high fidelity simulation model would provide a safe environment to teach thoracoscopic EA/TEF repair to novices. The study purpose was to evaluate validity evidence for performance measures on an EA/TEF simulator.

Methods: IRB-exempt data were collected from 12 self-reported “novice” and 8 “experienced” pediatric surgeons. Participants evaluated the EA/TEF repair simulator using survey ratings that were analyzed for test content validity evidence. Additionally, deidentified operative performances were videotaped and independently rated by two surgeons using the Objective Structured Assessment for Technical Skills (OSATS) instrument. Novice and experienced OSATS were compared with p < .05 significant.

Results: Participants had high overall simulator ratings. Internal structure was supported by high interitem consistency (α = .95 and .96) and interrater agreement (ICC [.52, .84] for OSATS ratings. Experienced surgeons performed at a significantly higher level than novices for all five primary and two supplemental OSATS items (p < .05).

Conclusion: Favorable participant ratings indicate the simulator is relevant to clinical practice and valuable as a learning tool. Further, performance ratings can discriminate experienced and novice performances of EA/TEF repair. These findings support the use of the simulator for performance assessment, representing the first validated measures from a simulator intended for pediatric surgical training.

⁎ Corresponding author. Ann and Robert H. Lurie Children’s Hospital of Chicago, Division of Pediatric Surgery, 225 East Chicago Avenue, Box 63, Chicago IL 60611–2605.
Tel.: +1 312 227 4210; fax: +1 312 227 9678.
E-mail address: kbarsness@luriechildrens.org (K.A. Barsness).

© 2014 Elsevier Inc. All rights reserved.

1. Background

Thoracoscopic esophageal atresia/tracheoesophageal fistula (EA/TEF) repair is an advanced minimally invasive procedure. Previous authors have described many “technical hurdles” [1], a “demanding” [2,3] operative environment and “significant challenges” [4] when introducing thoracoscopic EA/TEF into individual or group practices. Additionally, nearly all authors note that a learning curve exists for the skills necessary to safely complete a thoracoscopic EA/TEF repair. Traditionally, skill acquisition to overcome a learning curve occurs in the operating room. However, advances in simulation education may provide opportunities to move the learning curve outside of the operating room, potentially decreasing patient morbidity associated with the introduction of these new procedures.

Surgical simulation seeks to realistically provide a relevant operative experience in which the participant can safely make mistakes and receive feedback, to ultimately improve technical and cognitive skills [5]. Despite high perceived need and desire among trainees and educators alike, there are few opportunities for relevant simulation experiences in pediatric surgery [6].

We have previously presented our work on a prototype of a novel high fidelity thoracoscopic esophageal atresia/tracheoesophageal fistula (EA/TEF) repair simulator [7]. Based on early data analyses, there was a high perceived value of our simulator. The same work also offered preliminary validity evidence relevant to test content and internal structure, but indicated a need for refinements to the simulator and a deeper evaluation of validity evidence to fully support the use of the new model in simulation-based education in pediatric surgery.

We used this study to evaluate additional validity evidence relevant to test content and internal structure of the measures captured while performing a thoracoscopic esophageal atresia/tracheoesophageal fistula (EA/TEF) repair on a refined simulator. Further, we evaluated validity evidence relevant to relationships to other variables by testing our hypothesis that experienced minimally invasive pediatric surgeons would perform at a higher level than novice pediatric surgeons.
2. Methods

2.1. Study setting

After review and exempt approval by Ann and Robert H. Lurie Children’s Hospital of Chicago Institutional Review Board, data were collected during a national pediatric surgery meeting. Twenty pediatric surgeons (in-training or active practice) participated and were categorized as “experienced” or “novice”, based on self-reported experience with thoracoscopic EA/TEF repair. Eight experienced surgeons self-reported a mean of 27 (range 6–60) total EA/TEF repair procedures, and a mean of 11 (range 3–25) thoracic EA/TEF repairs. Twelve novice surgeons self-reported a mean of 21 (range 0–100) total EA/TEF repair procedures, and no more than 1 thoracoscopic EA/TEF repair. Experienced surgeons reported a mean lowest infant weight of 1.6 kg (range 0.6–2.0 kg) while novices reported a mean lowest infant weight of 5.7 kg (range 1.2–10.0 kg).

2.2. Simulator and data collection

As previously described [8], the EA/TEF repair simulator was assembled using a term neonatal rib cage (right side only), a stabilizing base and synthetic skin overlay. The simulator was completed with second trimester fetal bovine tissue that has been surgically modified into the classic C type anatomy, proximal esophageal atresia with distal tracheoesophageal fistula. Modifications of the originally described simulator included improvements to the scapular location and more consistent location of the carina/fistula at the origin of the 5th rib. All participants were shown a pictorial display of the specific procedural steps (Fig. 1). Participants were provided with 3 mm instruments and a 4 mm telescope (Karl Storz Endoscopy-America, Segundo, CA). Laparoscopic clip appliers and precut 5–0 silk sutures (Covidien, Mansfield, MA) were also supplied. Start time was the time of first incision. End time was removal of all instruments at end of procedure or 45 minutes, whichever occurred sooner.

2.3. Measures and rating procedures

There were two rating instruments used in this study. The first instrument was a self-report survey, completed by participants following their experience with the simulator. The 24-item survey consisted of 23 five-point rating scales measuring 6 domains (Physical attributes, Realism of materials, Realism of experience, Ability to perform task, Value, and Relevance) and one 4-point global rating to measure participants’ overall impression of the simulator. The second instrument consisted of nine items and was used to independently assess participants’ deidentified videotaped performances of EA/TEF repair by two faculty surgeons (KB/AC). The first five items of the instrument were from the previously validated Objective Structured Assessment for Technical Skills (OSATS) [9]. Our instrument included items that captured measures across five separate domains (Respect for tissue, Time and motion, Instrument handling, Flow of operation, and Knowledge of specific procedures) and were rated on a 5-point scale. Three items were intended to capture procedure-specific measures. These included two dichotomously scored items (Anastomosis patent, Anastomosis completed), and one continuously scored item (Total time to complete task).

2.4. Analyses

To analyze evidence relevant to test content we employed a Rasch model. Analysis was performed using the Facets software v. 3.68 (Linacre 2011, Chicago, IL). To evaluate content validity of the subjective measures from the self-report survey, we evaluated two Rasch indices—observed averages, and point-measure correlation statistics [10]. Content validity of the performance measures from the modified OSATS was similarly evaluated.

To evaluate evidence of internal structure for the self-report survey, we estimated interitem consistency using Cronbach alpha. For the modified OSATS instrument we estimated both interitem consistency (Cronbach alpha), and interrater agreement of the two faculty OSATS ratings using average measures of a two-way random intraclass correlation (ICC).

---

Fig. 1. Procedural steps for the simulated thoracoscopic EA/TEF repair.
To evaluate validity evidence relevant to relationships to other variables, specifically, discriminant validity, we compared the means of novice and experienced MIS OSATS ratings using a one-way analysis of variance (ANOVA). The summed OSATS scores were also correlated with the summed scores of the four supplemental items using Pearson correlation. We considered \( p < .05 \) to be statistically significant.

### 3. Results

#### 3.1. Evidence relevant to test content

##### 3.1.1. Observed averages

For the survey items, overall observed averages for all participants were high (4.5 on a 5 point scale). In descending order, the combined observed averages of the six domains were 4.7 (Relevance to practice), 4.7 (Value), 4.7 (Physical attributes), 4.7 (Realism of materials), 4.5 (Realism of experience), and 4.3 (Ability to perform task). Closer examination indicated the highest-rated items from the survey were Physical attributes-intercostal space (4.8), Relevance of materials-ribs (4.8), Physical attributes-Landmark tactility, scapula (4.7), and Relevance of experience—realism of location of “fistula” (4.7), while the lowest ratings were associated with Ability to properly place trocars (4.3), Realism of experience—amount of pressure needed to place trocars into thoracic space (4.1), and Ability to repair esophageal atresia (3.8). The observed average of Global opinion ratings was 3.3 (out of 4.0), indicating that on average, participants believed the new thoracoscopic EA/TEF repair simulator could be considered for training, but could be improved slightly. The primary write-in suggestion for simulator improvement was an esophageal tube to assist in identifying the proximal esophagus.

##### 3.1.2. Point-measure correlations

For the survey, all 24 items had positive point-measure correlations, ranging between 0.28 and 0.63. For the modified OSATS instrument, all five items also had positive point-measure correlations, ranging between 0.35 and 0.72 (Table 1).

For the purpose of this study, high observed-averages from the survey suggest high perceived value for the simulator, while positive point-measure correlations for the survey and modified OSATS attest to the “psychometric soundness” of the items in each instrument. These data support the assumption that participants’ ratings reflect the intended concepts—perceived value of the simulator and quality of performance during an EA/TEF repair. These findings support validity evidence relevant to test content.

#### 3.2. Evidence relevant to internal structure

Interitem consistency of the 23 survey items for all participants was estimated to be high (\( \alpha = 0.89 \)). Interitem consistency of the five OSATS items for the two faculty raters was also high (\( \alpha = 0.95 \) and 0.96). Interrater agreement across the five OSATS items ranged from 0.52 to 0.84 (Table 2). High interitem consistency estimates for both the 23-item survey and the five-item OSATS suggested adequate item reliability (internal consistency) for both instruments, while high interrater agreement for the five OSATS items indicated the experienced surgeons’ performance ratings were consistent when using the OSATS instrument. These findings offer support of evidence relevant to internal structure, indicating that the items of both instruments are “psychometrically sound” and are being used with an adequate amount of reliability. These data allow us to make inferences from our findings with a high degree of confidence.

#### 3.3. Evidence relevant to relationship to other variables

One-way analysis of variance (ANOVA) comparing the means of novice and experienced surgeons OSATS cumulative ratings indicated a statistically significant difference [\( F(1,16) = 8.17, p = .01 \)], where experienced surgeons (observed average = 21 ± 2.97) performed at a significantly higher level than novices (observed average = 15 ± 4.5). As summarized in Table 2, experienced surgeons also performed at a significantly higher level than novices for each of the five OSATS items (\( p < .05 \)), and two of the three supplemental procedure-specific items (\( p < .05 \)). There was no significant difference found for Time to complete (\( p = .127 \)).

Additionally, to support validity evidence relevant to relationship to other variables, the correlation of averaged summed five OSATS ratings and summed three supplemental procedure-specific items was high, Pearson’s \( r(16) = 0.731 \), \( p = .001 \). These data suggest that OSATS and the three supplemental items could be considered useful measures to assess performance on the EA/TEF simulator.

### 4. Discussion

With increased interest in simulator development in medical education, a number of researchers have focused on the evaluation of validity evidence of measures from novel simulators. Traditionally, when validity evidence is reported, researchers turn to the Standards for Educational and Psychological Testing (Standards), the guide developed jointly by American Education Research Association (AERA), American Psychological Association, and the National Council on Measurement in Education [11]. The current Standards identified five different sources of validity evidence, a) test content, b) internal structure, c) response processes, d) relationships to other variables, and e) consequences of testing. We used this research to ensure validity evidence relevant to test content during simulator refinement, and to evaluate additional evidence relevant to internal structure and relationships to other variables. We sought to further refine our simulator, and to expand the depth and breadth of our evaluation, so that we might have more confidence in the inferences made relative to performance assessment on the simulator.

The original simulator was refined both in the physical structure of the rib cage (position and size of the scapula) and in the internal tissue (carina/fistula located at origin of 5th rib and consistent esophageal gap 15–20 mm). Following refinement, observed averages for Value, Physical attributes, and Realism of materials all increased from values

---

**Table 1**

Inter-item consistency and inter-rater agreement of OSATS items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Point-measure correlation</th>
<th>ICC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Respect for tissue</td>
<td>.72</td>
<td>.84</td>
<td>.57, .94</td>
</tr>
<tr>
<td>2. Time and motion</td>
<td>.61</td>
<td>.76</td>
<td>.36, .91</td>
</tr>
<tr>
<td>3. Instrument handling</td>
<td>.75</td>
<td>.73</td>
<td>.28, .90</td>
</tr>
<tr>
<td>4. Flow of operation</td>
<td>.71</td>
<td>.82</td>
<td>.51, .94</td>
</tr>
<tr>
<td>5. Knowledge of specific procedures</td>
<td>.35</td>
<td>.52</td>
<td>.29, .82</td>
</tr>
</tbody>
</table>

**Table 2**

Comparison of novice verses experienced surgeons’ observed averages.

<table>
<thead>
<tr>
<th>OSATS†</th>
<th>Novice (n = 12)</th>
<th>Experienced (n = 8)</th>
<th>F</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative score (range, 0–25)</td>
<td>15 ± 4.5</td>
<td>21 ± 2.97</td>
<td>8.17</td>
<td>.010</td>
</tr>
<tr>
<td>1. Respect for tissue</td>
<td>3.2 ± .88</td>
<td>4.1 ± .61</td>
<td>4.93</td>
<td>.040</td>
</tr>
<tr>
<td>2. Time and motion</td>
<td>2.6 ± .85</td>
<td>4.0 ± .65</td>
<td>14.89</td>
<td>.001</td>
</tr>
<tr>
<td>3. Instrument handling</td>
<td>2.8 ± .96</td>
<td>4.0 ± .85</td>
<td>8.83</td>
<td>.009</td>
</tr>
<tr>
<td>4. Flow of operation</td>
<td>3.4 ± 1.10</td>
<td>4.4 ± .69</td>
<td>4.53</td>
<td>.049</td>
</tr>
<tr>
<td>5. Knowledge of</td>
<td>3.4 ± 1.10</td>
<td>4.4 ± .61</td>
<td>6.12</td>
<td>.025</td>
</tr>
<tr>
<td>6. Patent anastomosis</td>
<td>.46 ± .52</td>
<td>1.0 ± 0.0</td>
<td>8.17</td>
<td>.009</td>
</tr>
<tr>
<td>7. Complete anastomosis</td>
<td>.46 ± .52</td>
<td>9 ± .35</td>
<td>8.17</td>
<td>.011</td>
</tr>
<tr>
<td>8. Time to complete (min)</td>
<td>34.67 ± 11.29</td>
<td>27.4 ± 6.02</td>
<td>2.55</td>
<td>.127</td>
</tr>
</tbody>
</table>

* Ratings were averaged for two raters.
noted with the first prototype [7]. Additionally, the observed average for the overall global rating was also improved from the original prototype. These changes cumulatively suggest that the simulator more realistically replicates a thoracoscopic EA/TEF repair than the original prototype. However, with a global rating of 3.3 (out of 4), there are still changes that can be made to further refine some of the procedure-specific details. The most consistently suggested refinement was the addition of an esophageal tube to aid in identifying the proximal pouch. This simple addition will be included in subsequent studies using this simulator.

We next sought to determine if the simulator reliably tests skill and whether we would be able to differentiate between levels of skill [12,13]. Because of sample size, we used simple dichotomous comparisons of experienced and novice pediatric surgeons. Although these levels were arbitrarily defined for the purposes of this study, we were able to clearly differentiate between the two groups when performance was assessed using the modified OSATS instrument. The only measure that did not reach statistical significance was Time to Complete. The study design was that all participants had a maximum of 45 minutes to complete all tasks. Even if all tasks were not completed, time was recorded to end at 45 minutes. Perhaps more telling than Time to Complete is that within the 45-minute time limit, all experience surgeons completed the anastomosis (last task), but less than half of the novice surgeons had a complete and/or patent anastomosis. Cumulatively, these performance data provide early validity evidence relevant to internal structure and relationships to other variables.

There are limitations related to the interpretation and applications of the findings from this study. First, the sample size was small. For the purpose of this study, the smaller sample size may have decreased the precision of the measures, which should be considered. Additionally, we used rather liberal definitions of “experienced,” requiring only three thoracoscopic and at least six total EA/TEF repairs. Few surgeons would define “experienced” so loosely. Yet, even with this rather liberal definition, there were clear technical differences noted between the two groups.

The last two measures of validity evidence to support the use of this simulator in pediatric surgical education (response to other variables and consequences of testing) are not included in this study. The response to other variables is essentially determining the effect of performing on the simulator relative to a specific outcome measure. Example outcome measures may be shorter operating times, decreased morbidity or shorter hospital stays. Given the rarity of these procedures, these data will be slow to acquire. However, outcome data after simulation training are critical to collect, as they could provide evidence that we can successfully move the inherent risks of a learning curve from the operating room to a simulation-based environment. The final measure, consequences of testing, is interesting in the context of recent changes to American Board of Surgery Certification for General Surgery [14]. For the first time ever, a validated measure of skill (Fundamentals of Laparoscopic Surgery) and documentation of technical proficiency in specified procedures are both now required for board eligibility. Although not yet defined, it is reasonable to assume that similar changes may be on the horizon for pediatric surgery. Currently, there are no requirements for validated measures of skill during the certification or maintenance of certification process for pediatric surgery. The introduction of such measures remains highly controversial, despite widespread support for simulation-based education from the American College of Surgeons and the American Society of Anesthesiologists [15,16]. Continued investment in pediatric-specific simulation research is critical to ensure that validated simulation-based assessments exist in anticipation of potential changes in certification and maintenance of certification processes.

In conclusion, through continued refinement, we have demonstrated the value of a high fidelity simulation model for thoracoscopic esophageal atresia/tracheoesophageal fistula repair. Both novice and expert participants performing on the simulator found it realistic and relevant for pediatric surgical training. The addition of validated measures of performance on the simulator provides critical benchmarking data. These data are relevant to future studies on the use of the simulator in defined curricula for pediatric surgical training.

References


Discussion

Discussant: ANDREAS MEIER (Syracuse, NY): This is a really nice education study. Congratulations. I have a comment and a question.

I think the question is already partially answered by your last slide. The one thing that you are currently missing is how does the performance relate to performance in the operating room? You are currently working with discriminative validity which is a good first step but you still have to get content validity which is basically showing what does this do to the performance in the operating room. One study you could consider doing is filming in the operating room and then basically see if there is a correlation between the scores on the simulator versus their scores in the operating room.

My comment, the subcommittee for simulation in this institution is currently working on a national simulation curriculum and this would be a nice tool to add to that.

Response: DR. KATHERINE BARNES: One of the ideals is that we actually can show an outcome measure such as improved patient safety, because that’s really what it is about. Why simulators? We’re simulating to be better surgeons so that we can take care of
our patients better. Of note, although I present data on a thoracoscopic repair, these models are equally translatable to open procedures. In fact if someone starts thoracoscopic and they get confused or can’t complete the procedure, they can convert the model to an open procedure with ease and it doesn’t require changing to a different simulator.

Patient outcomes. Really that would take a multi-institutional study. It’s going to be very difficult when we are aiming this to be an educational tool, at least initially for fellows, because the fellow never does the entire operation. That is going to be a little bit difficult, but at a multi-institutional level we are looking at different measures to actually look at the outcomes related to this effort.

*Discussant: David Bliss (Los Angeles, CA):* My question for you is, there are some emerging data about surgical warm-up being useful for complex operations, meaning for things that we don’t do very often or that really tax the surgeon, and I would say this is an example of one, that a period of time spent right before the operation practicing these moves can improve performance across a lot of different measures. Have you looked at that and/or do you have any plans to look at that for this device?

*Response: Dr. Katherine Barsness:* The answer is yes and no both. This device is made with bovine tissue which no hospital is going to want next to the operating room, so for the true “just in time,” which means 10 minutes before you go into the OR or 30 minutes before, that’s going to be a little bit difficult with this current high fidelity procedure. What my model tests is not just technical skill. My model actually tests cognitive skill as well when it’s the high fidelity, but for a middle fidelity model that we are also working on and should have available out by the next meeting, will be made of synthetic tissue that can just work on the tasks that “just in time” is aiming for so, yes, we will have those available in the future. We are just not quite there yet. We have been focusing on the high fidelity model but this version, is not ready for use near the OR because of the bovine tissue issue.

*Unidentified Speaker (moderator):* I have one quick question and then we have to end the discussion. I want to congratulate you and I know I speak for a lot of our colleagues. You catapulted us from a simulation standpoint, but I’ve watched you slave over these things, how much time you’ve put in, sewing these anastomoses. How do you scale this, because it’s you putting it together? How does everyone else in the country do this now?

*Response: Dr. Katherine Barsness:* Right now I receive tissue blocks from a butcher house in Lubbock, Texas, so I am at their mercy. I’ll say many of the specimens they sent for the meeting and for the course this week were suboptimal. Really what it involves is – I live in Chicago and we are known for steak, so what it involves is me actually working with local butcher shops where you can’t thaw tissue and then re-freeze and then thaw again. You have to just freeze it once and then use it, so it involves actually doing the processing at the time the fetal calf, or actually its mother, is going to slaughter. It involves actually getting to the tissue at that point, making the modifications and then freezing it at that point. That has several advantages. One is, it is quickly assembled, but also it has the advantage of diminishing the size of the specimen so that there are lower shipping costs which are all weight based. If we can eliminate the extra liver, eliminate the extra intestine, the extra this and that, it also helps.