JOURNAL OF LAPAROENDOSCOPIC & ADVANCED SURGICAL TECHNIQUES
Volume 25, Number 3, 2015

DOI: 10.1089/lap.2014.0358

Evaluation of Three Sources of Validity Evidence for a Laparoscopic Duodenal Atresia Repair Simulator

Katherine A. Barsness, MD,[1,2] Deborah M. Rooney, PhD,[3] Lauren M. Davis, BA,[4] and Ellie O'Brien, BS[4]

Abstract

Purpose: Laparoscopic duodenal atresia (DA) repair is a relatively uncommon pediatric operation requiring advanced minimally invasive skills. Currently, there are no commercial simulators available that address surgeons' needs for refining skills associated with this procedure. The purposes of this study were (1) to create an anatomically correct, size-relevant model and (2) to evaluate the content validity of the simulator.
Materials and Methods: Radiologic images were used to create an abdominal domain consistent with a full-term infant. Fetal bovine tissue was used to complete the simulator. Following Institutional Review Board exempt determination, 18 participants performed the simulated laparoscopic DA repair. Participants completed a self-report, six-domain, 24-item instrument consisting of 4-point rating scales (from 1 = not realistic to 4 = highly realistic). Validity evidence relevant to test content and response processes was evaluated using the many-facet Rasch model, and evidence of internal structure (inter-item consistency) was estimated using Cronbach's alpha.
Results: The highest observed averages were for ''Value as a training and testing tool'' (both observed averages = 3.9), whereas the lowest ratings were ''Palpation of liver'' (observed average = 3.3) and ''Realism of skin'' (observed average = 3.2). The Global opinion rating was 3.2, indicating the simulator can be considered for use as is, but could be improved slightly. Inter-item consistency was high ($\alpha = 0.89$).
Conclusions: We have successfully created a size-appropriate laparoscopic DA simulator. Participants agreed that the simulator was relevant and valuable as a learning/testing tool. Prior to implementing this simulator as a training tool, minor improvements should be made, with subsequent evaluation of additional validation evidence.

Introduction

Laparoscopic duodenal atresia (DA) repair is a complex and technically challenging operation. It also has a steep learning curve, as evidenced by higher conversion rates and operative times for surgeons with little or no experience in the laparoscopic approach to DA repair.[1] However, more concerning than the duration of the operation is that early in the evolution of the technique, there were several anecdotal (unpublished) reports of high leak rates. In fact, one center completely abandoned the operation after an unacceptably high leak rate, only to resume the operation several years later, after substantial improvements in technique and skill.[2] These data cumulatively support not only a steep learning curve, but also that the learning curve may be placing infants at risk of serious morbidity.

Simulation-based education is a valuable adjunct to traditional surgical training. It is perfectly suited to maximize patient safety, yet provide ample opportunities for deliberate practice of complex technical skills. We have previously presented our work on the creation of several different minimally invasive simulators for congenital anomalies.[3–6] Using similar methods, we created a laparoscopic DA repair simulator. The purpose of this study was to evaluate three levels of validity evidence—test content, response processes, and internal structure—to support or refute its use in pediatric surgical education.

Materials and Methods

Study setting and participants

After review and exempt determination by Ann and Robert H. Lurie Children's Hospital of Chicago Institutional Review Board, data were collected during a national pediatric surgery meeting. In total, 18 participants contributed to this study. Ten

[1]Division of Pediatric Surgery, Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois.
[2]Departments of Surgery and Medical Education, Northwestern University Feinberg School of Medicine, Chicago, Illinois.
[3]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan.
[4]Innovations Laboratory, Northwestern Simulation, Center for Education in Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

participants completed the procedure and rated the simulator during an advanced neonatal minimally invasive surgery training course, whereas the remaining eight participants completed the procedure and rated the simulator outside of the course. The majority of participants had extensive experience with operative repair of DA, self-reporting a mean of 25 (range, 0–100) previous open DA repairs, with only one participant reporting no prior history with the operation. For analyses, the participants were divided into novice and experienced groups based on self-reported experience with laparoscopic DA repair. Novice participants ($n = 12$) self-reported a mean of 0.5 (range, 0–3) prior laparoscopic DA repairs, with nine reporting no prior experience with the laparoscopic approach. Experienced surgeons ($n = 6$) self-reported a mean of 8 (range, 4–15) prior laparoscopic DA repairs.

Simulator

As previously described, the external surround of the DA repair simulator was assembled using the lower half of a neonatal rib cage, a pelvis, and a stabilizing base (Fig. 1), with a synthetic skin overlay.[7,8] The simulator was completed with second-trimester fetal bovine tissue (Animal Technologies, Lubbock, TX). The abdominal block of tissue (spleen, four-compartment stomach, duodenum, small and large intestine, liver, and pancreas) was surgically modified until it was appropriately size for the abdominal space of the simulator. The tissue was then secured in a configuration consistent with a type I DA. Participants were provided with 3-mm instruments and a 4-mm telescope (Karl Storz Endoscopy–America, Segundo, CA) to complete the procedure.

Measures and rating procedures

All participants completed a 24-item, self-report, six-domain instrument consisting of 4-point rating scales (from 1 = not realistic to 4 = highly realistic), with a ''Not sure'' option recoded as missing data. Twenty-three items covered six domains including Physical Attributes, Realism of Materials, Realism of Experience, Ability to Perform Task, Value, and Relevance. Additionally, a 4-point global rating was used to measure participants' overall impression of the simulator.

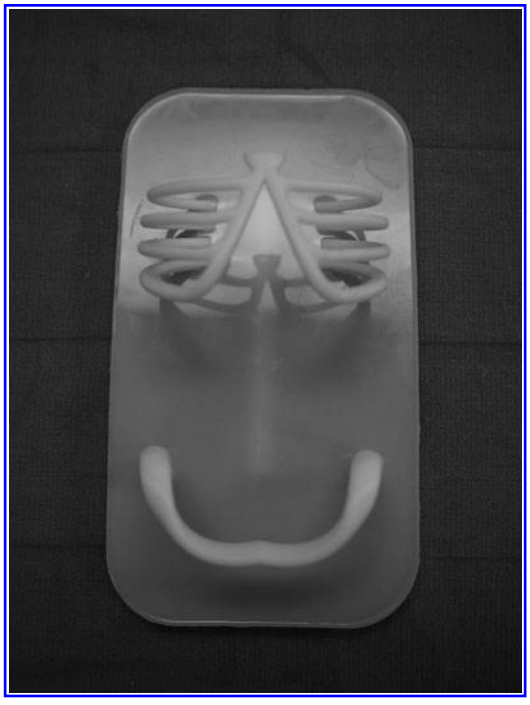

FIG. 1. Three-dimensional printed abdominal model used for laparoscopic duodenal atresia repair.

Analyses

In order to evaluate validity evidence, we used the Standards for Educational and Psychological Testing (Standards), the guide developed jointly by the American Education Research Association, the American Psychological Association, and the National Council on Measurement in Education.[9] The current Standards framework 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 work to evaluate three sources of validity evidence—test content, response processes, and internal structure.

To analyze the difference sources of validity evidence, we used methods from both modern measurement and classical test theories. Similar to methods used in previous work to evaluate evidence of test content, we used a many-facet Rasch model[10] to analyze three Rasch indices—observed averages, point-measure correlation, and Rasch item-fit statistic. To evaluate validity evidence relevant to response processes, we examined Rasch person fit statistics and rating differences across participants' experience levels using a many-facet Rasch model. Analyses of the self-report survey measures were performed using Facets software version 3.68.2 (Linacre, 2011).[11] To evaluate evidence relevant to internal structure we estimated inter-item consistency using Cronbach's alpha. Statistical analysis was performed using IBM SPSS statistical software (version 22.0; IBM Corp., Armonk, NY).

Results

Evidence relevant to test content

Observed averages. In descending order, the combined observed averages of the six domains were 3.9 (Value), 3.9 (Relevance), 3.6 (Realism of Experience), 3.5 (Realism of Materials), 3.5 (Physical Attributes), and 3.5 (Ability to Perform Task), out of a maximum score of 4.0. As shown in Table 1, closer examination indicated the highest-rated items from the survey were ''Value of the simulator as a training tool'' (3.9), ''Relevance to my practice'' (3.9), and ''Value of the simulator as a testing tool'' (3.8), whereas the lowest ratings were associated with ''Realism of skin'' (2.2), ''Realism of duodenal anatomy'' (2.4), and ''Realism—overall impression of the simulator materials'' (2.5). The observed average of the Global opinion ratings was 3.2 (out of 4.0), indicating that on average, participants believed the simulator ''could be considered for training, but could be improved slightly.''

Point-measure correlations. For the survey, all of the 24 items had positive point-measure correlations (range, 0.24–0.76). This indicates that each item of the survey contributed useful information to the construct as a whole. For the purpose of this work, positive point-measure correlations offer evidence of the raters' scores aligning with their observations, so that we can make inferences about the quality of the simulator with confidence.

Table 1. Observed Averages Across Six Domains, 24 Items

Domain	*Item*	*Rating (out of 4)*
Physical Attributes (average = 3.5)[a]	1. Abdominal length	3.72
	2. Abdominal width	3.72
	3. Total abdominal space	3.25
	4. Ability to palpate the liver edge	3.78
	5. Scaphoid appearance of the abdomen	3.61
	6. Overall position of structures inside the abdominal domain	3.65
Realism of Materials (average = 3.5)[a]	7. Skin	2.23
	8. Ribs/pelvis	3.44
	9. Stomach and duodenal tissue	2.67
	10. Overall impression—all simulator materials	2.56
Realism of Experience (average = 3.6)[b]	11. Amount of instrument resistance of the abdominal wall	3.29
	12. Realism of stomach anatomy during DA repair	3.67
	13. Realism of duodenal anatomy during DA repair	2.44
	14. Realism of other abdominal organs (liver, intestine)	3.47
	15. Does the simulator represent the expected experience during a neonatal DA repair?	3.69
Ability to Perform Tasks (average = 3.5)[c]	16. Acquisition of target trocar sites	3.72
	17. Ability to safely place trocars/instruments into abdominal cavity	3.72
	18. Ability to Kocherize the duodenum	3.50
	19. Ability to create duodenotomies	3.29
	20. Ability to complete duodenoduodenostomy	3.11
Value (average = 3.9)[d]	21. Value as training tool	3.94
	22. Value as testing tool	3.83
Relevance (average = 3.9)[e]	23. Relevance to practice	3.89
Overall rating	24. Global assessment	3.20

[a]Where 3 = adequate realism as is, but could be improved.
[b]Where 4 = highly realistic, no changes needed.
[c]Where 3 = somewhat easy to perform.
[d]Where 4 = great deal of value.
[e]Where 4 = highly relevant to my practice.
DA, duodenal atresia.

Rasch item-fit indices. All items' Rasch Outfit Mean-Square (MS) values fell between −2.0 and 2.0, suggesting a reasonable amount of variability in responses (agreement) for all items. Five items' Outfit MS values fell below 0.5, indicating a relatively high degree of agreement. For the purpose of this work, this finding supports validity evidence relevant to test content when paired with high observed averages. These items—5 (Scaphoid appearance of the abdomen), 10 (Overall impression—all simulator materials), 14 (Realism of other abdominal organs—liver, intestine), 21 (Value as training tool), and 22 (Value as testing tool)—had item Outfit MS values that ranged between 0.27 and 0.43 and high observed averages that ranged from 3.5 to 3.9. One item, item 24 (Overall global), had an Outfit MS value of 1.55, indicating slightly elevated variability in ratings. Deeper examination of the ratings for this particular item indicated a single rating of 1 (''This simulator requires a number of improvements before it can be considered for use in neonatal DA repair training''). Given the small sample size, it is likely that the single extreme rating affected the Outfit MS value for this item.

Evidence relevant to response processes

Rasch person fit indices. Beginning with the underfitting participants, none of the 18 participants had outfit MS statistics higher than the ideal (1.5), indicating a reasonable amount of ratings variability (error). Although not problematic for the purpose of our study, a review of overfitting responses indicated two (11.1%) participants had Outfit statistics below the acceptable boundary of 0.5, suggesting a relatively high degree of consistency in these participants' independent ratings. Closer examination of these participants' ratings indicate that there may have been a ceiling effect, with ratings of 3.7 (standard deviation = 0.42) and 3.9 (standard deviation = 0.21). Analysis indicated the majority (88.9%) of participants had Outfit statistics within acceptable limits, suggesting inferences from this group of participants aligned with the ''average'' participant, further supporting evidence relevant to response processes.

Rating difference across novice and experienced participants. Experienced and novice surgeons had identical overall ratings (observed average = 3.5) ($P = .75$). Item-level examination of rating differences between experienced and novice surgeons using a Mann–Whitney U test indicated no statistical differences for any of the 14 items ($P > .05$).

Evidence relevant to internal structure

Inter-item consistency of all items was estimated to be high ($\alpha = 0.89$). This index offers a measure of control and when adequately high indicates these assessment items are grouped appropriately and measure the same general construct. These data allow us to make inferences from our findings with a high degree of confidence and offer evidence of internal structure.

Discussion

Within pediatric hospitals, preventable adverse events most commonly occur in surgical patients, with infants accounting for a staggering 50% of these errors.[12] Although not all of these errors are directly attributable to new technology, learning curves associated with the application of new technologies, devices, and/or procedures remain a preventable source of perioperative error. As a key component of risk mitigation strategies, simulation-based educational tools are being sought across several different surgical disciplines and subspecialties. A simulated operating room provides for maximal patient safety, while still offering opportunities for learners to practice to proficiency. For pediatric surgeons, a limitation in the use of simulation-based education has been the lack of relevant simulation models for complex operations. To this end, we sought to create a realistic and relevant laparoscopic DA repair simulation and to evaluate three levels of validity evidence—test content, response processes, and internal structure—to support or refute its use in pediatric surgical education.

Our initial validity evidence suggests that our novel laparoscopic DA repair simulator is valuable as an educational and testing tool, is relevant to clinical practice, and has many of the physical attributes of an infant with DA. These findings are supported by high observed averages across all domains, high estimated internal consistency across all quality measures of the simulator, and high participant agreement indicated by the Rasch person and fit indices. These results support validity evidence relevant to test content, response processes, and internal structure, as defined by the Standards for Educational and Psychological Testing.[9]

The highest ratings across all 24 items were ''Value as a training tool,'' ''Value as a testing tool,'' and ''Relevance to practice.'' Not only were these the highest rated categories, but these ratings were reflective of a fairly diverse pool of participants with variable experience in both open and laparoscopic repair of DA. These data support the use of this model as an educational tool not only for graduate medical education, but also for continuing medical education. However, these data are perhaps even more interesting given that the lowest scores were in ''Realism of materials'' and ''Realism of duodenal anatomy.'' Perhaps the participants' evaluation of the model was more a reflection of its future potential, rather than in its current form.

Low scores for ''Realism of materials'' and ''Realism of experience'' are likely both rooted in the same design flaws of the model. The preparation of the fetal bovine tissue was difficult and labor intensive, especially in the beginning of our research. The majority of the liver must be removed and then stabilized in the right upper quadrant. At the time of data collection, we had not yet developed a reliable method of liver stabilization, which in turn leads to excess movement of the liver and duodenum during the procedure. We have since developed a liver stabilizer to help with fixation of the liver in an anatomically correct location and orientation.

The bovine stomachs present a unique opportunity, while, at the same time, a challenging anatomic configuration problem. The fetal bovine stomach has four different compartments that are all interconnected. Relevant to DA, the transition from the fourth stomach to the proximal duodenum is notable for having a marked size discrepancy. For all practical purposes, the transition looks like a type I DA (Fig. 2). The transition of the third stomach into the fourth stomach then mimics the appearance of the pylorus. Although the pure serendipity of such a finding is wonderful, the accurate positioning of the stomachs and the duodenum proved to be difficult. Again, at the time of data collection, we were early in our own learning curve for the preparations of the specimens. With added experience and additional surgical modifications of all of the stomach compartments, the anatomy in subsequent versions of the model is more similar to the findings in infants with DA. The tissue stabilization and anatomic modifications continue to evolve as we collect validity evidence with subsequent iterations of the model.

FIG. 2. Bovine anatomy replicating type I duodenal atresia.

Although not explicitly queried on the survey, several participants also noted that the mucosa and serosa were very difficult to differentiate from each other. Specifically, with the lack of tissue perfusion, both layers of the intestine were the exact same pale flesh-like color. Newer versions of the model include stomach and duodenal luminal flushes with red food coloring, resulting in a pink-red enhancement of the mucosa without affecting the color of the serosa.

We have addressed all of the apparent design flaws with significant structural modifications, and the participant survey has been modified to allow continued evaluation of these critical quality measures of the simulator. Yet, despite these flaws, participants overwhelmingly support the use of the model as an educational tool with the extremely high ''Value'' and ''Relevance'' ratings. Finally, the overall ''Global assessment'' rating was 3.2, consistent with ''this simulator can be considered for use in laparoscopic DA training, but could be improved slightly.'' These data support the use of a modified version of the laparoscopic DA repair simulator in pediatric surgical education.

The biggest question—whether deliberate practice on our laparoscopic DA simulator will be able to improve patient outcomes—remains to be studied. Although not surprising given the relative paucity of pediatric-specific simulation devices, there are no data on the transferability of pediatric-specific surgical skills from the simulation lab to the operating room. There are several barriers limiting our ability to collect these data. First, the design modifications for the model need to be further evaluated, ensuring that any new modifications result in improved realism. Only with these data can we begin to collect performance metrics from a variety of different learner groups. Second, we have not yet evaluated whether performance metrics on the simulator can objectively discriminate

between novice and experienced pediatric surgeons. Third, a comprehensive curriculum will need to be created that incorporates all of the cognitive, technical, and nontechnical skills integral to the safe performance of a duodenoduodenotomy. After performance improvements are documented in simulation, we can begin to query the translation of these results to our patients. Yet, the final barrier is one that is inherent in the majority of operations for congenital anomalies—they are rare. It will only be possible to examine patient-level data with a multi-institutional trial.

There are several limitations related to the interpretation and applications of the findings presented in this study. These data were collected from a volunteer pool of participants at a national pediatric surgery meeting. We are unable to determine if the baseline opinions, interests, and/or performance characteristics of volunteers are different from a cross-section of pediatric surgeons across a variety of regions and practice models. Second, our participation pool is relatively small, which may impact the variability of the ratings, thereby falsely reassuring us as to the validity of our measures. Third, these data represent the first prototype of a simulation model that has undergone several subsequent changes. Although we anticipate the changes will only strengthen our evidence, ongoing evaluation of validity evidence is necessary. Finally, we have only begun to examine three of the five levels of validity evidence, and ongoing evaluation is required to ensure optimum simulator quality.

In conclusion, we have created a highly valued and relevant laparoscopic DA repair simulator. Initial validity evidence relevant to test content, response processes, and internal structure indicates structural refinement requirements, particularly as it relates to the tissue preparation and positioning of the tissue within the model, and the ongoing collection of additional validity evidence from a refined model.

Acknowledgments

The authors would like to thank Northwestern Simulation at Northwestern University Feinberg School of Medicine for the continued support of our research. We would also like to thank David Irvin, Manager of Simulation Operations, Northwestern Simulation, for his never-ending enthusiasm and commitment to the success of our educational research. Finally, we would like to thank Karl Storz Endoscopy–America for their unwavering support of pediatric surgical education.

Disclosure Statement

No competing financial interests exist.

References

1. Jensen AR, et al. Laparoscopic versus open treatment of congenital duodenal obstruction: Multicenter short-term outcomes analysis. J Laparoendosc Adv Surg Tech A 2013;23:876–880.
2. van der Zee DC. Laparoscopic repair of duodenal atresia: Revisited. World J Surg 2011;35:1781–1784.
3. Barsness KA, Rooney DM, Davis LM. Collaboration in simulation: The development and initial validation of a novel thoracoscopic neonatal simulator. J Pediatr Surg 2013;48: 1232–1238.
4. Barsness KA, Rooney DM, Davis LM. The development and evaluation of a novel thoracoscopic diaphragmatic hernia repair simulator. J Laparoendosc Adv Surg Tech A 2013;23:714–718.
5. Davis LM, Barsness KA, Rooney DM. Design and development of a novel thoracoscopic tracheoesophageal fistula repair simulator. Stud Health Technol Inform 2013;184: 114–116.
6. Davis LM, Hawkinson EK, Barsness KA. The evolution of design: A novel thoracoscopic diaphragmatic hernia repair simulator. Stud Health Technol Inform 2014;196:89–95.
7. Hawkinson EK, Davis LM, Barsness KA. Design and development of low-cost tissue replicas for simulation of rare neonatal congenital defects. Stud Health Technol Inform 2014;196:159–162.
8. Hawkinson EK, Davis LM, Barsness KA. Design and development of a laparoscopic gastrostomy tube placement simulator. Stud Health Technol Inform 2014;196:155–158.
9. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association, 1999.
10. Wolfe EW, Smith EV Jr. Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. J Appl Meas 2007;8:243–290.
11. Lincare J. A user's guide to facets. Chicago: Winsteps, 2010.
12. Matlow AG, et al. Adverse events among children in Canadian hospitals: The Canadian Paediatric Adverse Events Study. CMAJ 2012;184:E709–E718.

Address correspondence to:
Katherine A. Barsness, MD
Division of Pediatric Surgery
Ann and Robert H. Lurie Children's Hospital of Chicago
Box 63
225 East Chicago Avenue
Chicago, IL 60611

E-mail: kbarsness@luriechildrens.org

This article has been cited by:

1. Hsiung Grace E., Schwab Ben, O'Brien Ellen K., Gause Colin D., Hebal Ferdynand, Barsness Katherine A., Rooney Deborah M.. 2017. Preliminary Evaluation of a Novel Rigid Bronchoscopy Simulator. *Journal of Laparoendoscopic & Advanced Surgical Techniques* **27**:7, 737-743. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
2. Takahiro Jimbo, Satoshi Ieiri, Satoshi Obata, Munenori Uemura, Ryota Souzaki, Noriyuki Matsuoka, Tamotsu Katayama, Kouji Masumoto, Makoto Hashizume, Tomoaki Taguchi. 2017. A new innovative laparoscopic fundoplication training simulator with a surgical skill validation system. *Surgical Endoscopy* **31**:4, 1688-1696. [CrossRef]
3. David C. van der Zee. 2017. Endoscopic surgery in children – the challenge goes on. *Journal of Pediatric Surgery* **52**:2, 207-210. [CrossRef]
4. Deie Kyoichi, Ishimaru Tetsuya, Takazawa Shinya, Harada Kanako, Sugita Naohiko, Mitsuishi Mamoru, Fujishiro Jun, Iwanaka Tadashi. 2017. Preliminary Study of Video-Based Pediatric Endoscopic Surgical Skill Assessment Using a Neonatal Esophageal Atresia/Tracheoesophageal Fistula Model. *Journal of Laparoendoscopic & Advanced Surgical Techniques* **27**:1, 76-81. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
5. Takazawa Shinya, Ishimaru Tetsuya, Harada Kanako, Deie Kyoichi, Fujishiro Jun, Sugita Naohiko, Mitsuishi Mamoru, Iwanaka Tadashi. 2016. Pediatric Thoracoscopic Surgical Simulation Using a Rapid-Prototyped Chest Model and Motion Sensors Can Better Identify Skilled Surgeons Than a Conventional Box Trainer. *Journal of Laparoendoscopic & Advanced Surgical Techniques* **26**:9, 740-747. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
6. Gause Colin D., Hsiung Grace, Schwab Ben, Clifton Matthew, Harmon Carroll M., Barsness Katherine A.. 2016. Advances in Pediatric Surgical Education: A Critical Appraisal of Two Consecutive Minimally Invasive Pediatric Surgery Training Courses. *Journal of Laparoendoscopic & Advanced Surgical Techniques* **26**:8, 663-670. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
7. Satoshi Ieiri, Takahiro Jimbo, Yuta Koreeda, Satoshi Obata, Munenori Uemura, Ryota Souzaki, Yo Kobayashi, Masakatsu G. Fujie, Makoto Hashizume, Tomoaki Taguchi. 2015. The effect of forceps manipulation for expert pediatric surgeons using an endoscopic pseudo-viewpoint alternating system: the phenomenon of economical slow and fast performance in endoscopic surgery. *Pediatric Surgery International* **31**:10, 971-976. [CrossRef]
8. Katherine A. Barsness. 2015. Trends in technical and team simulations: Challenging the status Quo of surgical training. *Seminars in Pediatric Surgery* **24**:3, 130-133. [CrossRef]