Title:
The Hands and Head of a Surgeon: Modeling Operative Competency with Multimodal Epistemic Network Analysis

Authors:
A. R. Ruis[1,2]
(arruis@wisc.edu)

Alexandra A. Rosser[1,2]
(arosser@wisc.edu)

Cheyenne Quandt-Walle[2]
(cquandtwalle@wisc.edu)

Jay N. Nathwani[1]
(jnathwani@uwhealth.org)

David Williamson Shaffer[2]
(dws@education.wisc.edu)

Carla M. Pugh[1]
(pugh@surgery.wisc.edu)

Affiliations:
[1]University of Wisconsin–Madison, School of Medicine and Public Health, Department of Surgery
600 Highland Avenue
Madison, WI 53792

[2]University of Wisconsin–Madison, Wisconsin Center for Education Research
1025 West Johnson Street
Madison, Wisconsin 53706

Corresponding Author:
A. R. Ruis
(arruis@wisc.edu)

Abstract

Background

This paper explores a method for assessing intraoperative performance by modeling how surgeons *integrate* psychomotor, procedural, and cognitive skills to manage errors.

Methods

Audio-video data were collected from general surgery residents (N=45) performing a simulated laparoscopic ventral hernia repair. Errors were identified using a standard checklist, and speech was coded for elements related to error recognition and management. Epistemic network analysis (ENA) was used to model the integration of error management skills.

Results

There was no correlation between number or type of errors committed and operative outcome. However, ENA models showed significant differences in the integration of error management skills between high-performing and low-performing residents.

Conclusion

These results suggest that error checklists and surgeons' speech can be used to model the integration of psychomotor, procedural, and cognitive aspects of intraoperative performance. Moreover, ENA can identify and quantify this integration, providing insight on performance gaps in both individuals and populations.

Introduction

Medical errors are one of the leading causes of death in the United States.[1] However, research suggests that negative patient outcomes are predicted not by the number of errors committed but by *error management*: how well that surgeon identifies and recovers from operative errors.[2] Thus, there is a substantial need for interventions that help surgeons recover from errors, but the cognitive and decision-making processes which underlie surgeons' abilities to successfully manage their errors are complex, poorly understood, and rarely measured by standard assessments of surgical performance.

In a recent paper published in the *Annals of Surgery*, Madani and colleagues present a novel intraoperative performance framework which includes a wider (and more complex) set of surgical competencies determined to be essential for expert proficiency.[3] The framework, which is intended to provide a procedure-agnostic guide to performance assessment and curriculum design, characterizes surgical expertise as the integration of behaviors in five areas: psychomotor skills, declarative knowledge, interpersonal skills, personal resourcefulness, and advanced cognitive skills. In alignment with more general work on the development of expertise,[4–6] Madani and colleagues argue that mere possession of knowledge or mastery of individual skills in isolation is not sufficient; rather, expert surgeons must be able to integrate these and other competencies "efficiently, flexibly, and creatively" (p. 255) to achieve optimal patient outcomes.

Despite the importance of this integration, most assessment tools measure specific competencies only in isolation, often missing critical elements of operative ability.[7,8] For example, technical proficiency is typically evaluated with the Objective Structured Assessment of Technical Skills (OSATS),[9] while non-technical skills are often assessed using the Non-Technical Skills for Surgeons (NOTSS) rubric.[10] Though both purport to provide a global assessment of surgical competencies, neither measure takes into account how those competencies are effectively integrated to accomplish complex tasks. In addition, research suggests that such measures, while helpful for providing formative feedback, do not perform well as assessments of aptitude.[11–13] There are, of course, many other assessment tools widely used for evaluating and assessing surgical performance—including oral and written examinations, task-specific and global rating scales, final product analyses, and documentation of critical failures—yet most approaches that evaluate performance *in situ* are based on assessment of technical skill and simple metrics such as the time it takes to complete a procedure.[14]

Using the framework developed by Madani and colleagues as a guide, we conducted a study to explore a method for identifying, measuring, and visualizing how and to what extent general surgery residents were able to integrate different elements of intraoperative performance. To do this, we used a dataset previously reported on in the *American Journal of Surgery* by Law Forsyth and colleagues.[15] In particular, we explored how residents identified and managed errors while completing the final steps of a simulated laparoscopic ventral hernia repair. For this study, we used *epistemic network analysis* (ENA), a statistical technique for constructing dynamic network models that quantify and visualize the structure and strength of association among elements of complex task performance over time.[16–19] ENA has been used to model and assess the integration of behaviors during highly technical or complex problem-solving activities in a range of domains, including engineering,[20–22] urban planning,[23,24] investigative journalism,[25] primary care,[26] and trauma surgery.[27] ENA thus provides an objective, quantitative method for measuring the integration of surgical skills, knowledge, and decision-making in an authentic operative context.

Methods

Setting and participants

Participants ($N = 45$: 21 women, 24 men) were general surgery residents (PGY1−5) from seven different institutions. Residents performed the final steps of a laparoscopic ventral hernia (LVH) repair on a physical, box-style simulator designed to represent the abdominal cavity of a patient with a ventral

hernia.[28] All necessary open and laparoscopic equipment for a mesh repair was provided. Participants were informed that two anchoring sutures had already been brought through the patient's skin and were given fifteen minutes to complete the repair by retrieving and securing the last two sets of sutures and placing five tacks to secure the mesh to the abdominal wall. Participants were not given any information about the purpose of the study. Trained researchers acted as medical student-level assistants during the simulation and introduced themselves as such. Assistants were not permitted to explain any aspects of the LVH procedure to the participants. Participants were not trained on how to speak during the procedure beyond being instructed to verbalize their needs as acting surgeons to their assistants.

Data collected

We audio and video recorded all simulated procedures (laparoscopic and external video). For each participant, we transcribed the audio data—each participant's intraoperative speech, or *discourse*—into *utterances*, where an utterance was defined as a continuous turn of talk; a period of silence lasting longer than three seconds was used to mark the end of an utterance. All utterances were timestamped for integration with errors committed during the simulation. Discourse only included utterances exchanged between participants and assistants.

Outcome measure

A surgeon rater trained in grading simulated hernia skins graded all hernia repairs for completion and quality using a previously validated checklist, which includes both technical and cognitive performance measures.[15] Possible outcome scores range from 0 to a maximum of 24. Outcome scores for the participants in this study ($n = 45$) ranged between 3 and 24 ($M = 14.29$, $SD = 5.36$). Outcome scores were normally distributed according to a Shapiro-Wilk normality test ($W_{OS} = 0.97$, $p_{OS} = 0.29$).[a]

Error identification

We used a standard checklist to identify six discrete and unrelated errors that can occur during the latter half of an LVH repair (Table 1).[29] Each error was categorized as a *Cognitive*, *Visuospatial*, or *Motor* error following the framework of intraoperative performance developed by Madani and colleagues.[3] *Cognitive Errors* indicated problems in the advanced cognitive skill and declarative knowledge domains of Madani's framework, while *Visuospatial* and *Motor Errors* indicated problems in the psychomotor domain. Errors occurred naturally; that is, they resulted only from residents' decisions and actions. Errors committed during the simulated procedure were recorded and timestamped live by a trained rater and validated afterward by a second, independent rater using video data. Disagreements between the two raters were resolved through discussion and review of the video to produce a single set of ratings on which both raters agreed. Errors were then integrated chronologically into the transcripts of recorded discourse based on the times at which they occurred.

[a] The Shapiro-Wilk test assumes a normal distribution as its null hypothesis. Therefore, a p-value of > 0.05 indicates no statistical basis for rejecting the null hypothesis, and thus a high probability of normality.

Table 1: Error Descriptions

Error	Type of Error	Description & Adverse Events
Inserts Tool without Visualization	Cognitive	The surgeon inserts a laparoscopic instrument without camera visualization, which risks nicks and punctures to organs and blood vessels.
No Incision Prior to Endoclose	Cognitive	The surgeon fails to make an incision in the skin prior to inserting the endoclose. This failure can make insertion more difficult, thus inviting the use of more force than necessary, which can then lead to puncture of organs.
Grabs Extra Suture	Visuospatial	The surgeon pulls up two sutures with a suture passer at the same time. This error can result from a surgeon mistaking the location of the grasper and/or suture passer relative to the sutures, or it can result from a failure to visualize the structural consequences of grabbing two sutures at once.
Two Sutures Same Hole	Visuospatial	The surgeon pulls up a second suture through the same hole as the first, causing the mesh to fall back into the abdominal wall. Similar to *Grabs Extra Suture*, this error indicates that the surgeon has failed to visualize the structural consequences of pulling two sutures through the same hole.
Tacker Slips	Motor	The tacker slips on the mesh while the surgeon is attempting to place a tack. This error can pull the mesh if it has not already been secured, or it can cause unintended tack placement.
Drops Tool	Motor	The surgeon drops an instrument. This can contaminate the operative space and potentially injure the patient or operative staff.

Discourse coding

To identify elements of residents' speech associated with error recognition and management, we conducted a *grounded analysis*[30] of residents' discourse during the simulated procedure. In a grounded analysis, transcripts are evaluated qualitatively to determine the presence or lack of potentially meaningful patterns of behavior. Our analysis focused on behaviors related to error recognition and recovery, and we used the framework of Madani and colleagues[3] to inform this analysis. In many cases, we observed that residents who exhibited good error management began by recognizing an issue (often by expressing dismay or frustration) and/or identifying a problem. They subsequently developed a plan to address the issue and gave instructions to the assistant accordingly. While the sequence was not always the same, residents who managed their errors well typically identified the problem and verbalized a plan to address it. The following example illustrates this pattern:

Line 1 *Alright, now look down and find that one suture that's left.*
Line 2 *I'm going to grab both.*
[Error: Grabs Extra Suture]
Line 3 *See, I got to come under you a little bit. You can just kind of stay still and I'll do it.*
Line 4 *There we go.*
Line 5 *Did it fall down? That's fine, no worries.*
Line 6 *Can I get both at once? Hmm.*
Line 7 *Let me pull up one and then I'll get the other one and since I guess I can't grab them both.*

The resident commits an error (Line 2), recognizes the problem (Line 5), and then devises a plan to fix it (Line 7).

To operationalize these elements of error management, we created four discourse codes: *Frustration*, *Identifying Errors*, *Operative Planning*, and *Giving Instructions* (Table 2). Importantly, these four codes are *procedure-agnostic*: that is, each code identifies an interpersonal or cognitive attribute that is universal in surgical practice. Per the framework developed by Madani and colleagues, *Identifying Errors* and *Operative Planning* are advanced cognitive skills necessary for operative success, while *Giving Instructions* is a key interpersonal skill. *Frustration* is a basic element of personal resourcefulness (recognition and management of stress, attention, and operative goals).

We developed automated coding algorithms for each code. For example, to automate the code *Identifying Errors*, we developed an algorithm that identifies verbalizations of mistakes in the discourse. Regular expressions ensure accurate string matching. For instance, the regular expression /bfell/b identifies instances of "fell"—as in, "the suture fell"—but not words containing "fell", like "fellow".

All four automated coding algorithms were validated by two trained human raters. For each code, the human raters and the coding algorithm independently rated a random sample of 50 utterances. Cohen's kappa was calculated between the two human raters and between each human rater and the coding algorithm. To determine whether the kappa values obtained for these samples could be reasonably generalized to the whole dataset, Shaffer's rho (ρ) was calculated for each kappa using the rhoR package for the R statistical computing software platform.[31] Rho uses an empirical sampling process that produces, for any inter-rater reliability statistic, an estimate of the expected Type I error rate at a given sample size.[32] Because kappa was greater than or equal to 0.80 and rho was less than 0.05 for every code and all combinations of raters (Table 3), we used the automated coding algorithms to code all the utterances in the dataset prior to ENA analysis. This automation reduced the total number of utterances that would need to be coded by human raters from 3,194 to 50—a considerable reduction in labor.

Table 2: Discourse Codes

Code	Definition	Example	Human 1 vs. Human 2		Human 1 vs. Computer		Human 2 vs. Computer	
			Kappa	Rho	Kappa	Rho	Kappa	Rho
Frustration	Expressing exasperation, dismay, or frustration	"Ugh, this is not so easy."	1.00	0.02	0.90	0.04	0.90	0.01
Identifying Errors	Recognizing a problem or something that needs to be corrected	"See how the port doesn't make it through the abdominal wall? That's the problem."	0.90	0.01	0.85	<0.01	0.80	0.03
Operative Planning	Describing a plan or strategy on the fly or indicating what needs to happen next	"We need to insufflate the abdomen."	0.88	<0.01	0.92	0.01	0.82	<0.01
Giving Instructions	Giving an instruction to or making a request of the assistant	"Can you show me the port?"	0.87	0.01	0.88	0.02	0.88	0.02

Epistemic network analysis (ENA)

To model error management, we used ENA, a statistical modeling tool which is described in detail elsewhere.[16–19] To do this, ENA uses statistical and visualization techniques to identify, quantify, and represent connections among coded behaviors as network models. For example, if a surgeon often responds to grabbing an extra suture by expressing frustration, but seldom by developing an operative plan, the resulting ENA network of connections between that surgeon's errors and discourse will possess a stronger connection between *Grabs Extra Suture* and *Frustration* than between *Grabs Extra Suture* and *Operative Planning*. Based on the different connection strengths between the error (*Grabs Extra Suture*) and the two discourse codes (*Frustration* and *Operative Planning),* we can infer that this surgeon was more likely to respond to errors by expressing frustration and less likely to do so by verbalizing the steps that could be taken to manage them.

In this study, we assessed error management by modeling only those connections that occurred between errors and any of the four discourse codes that appeared in the utterances that immediately followed an error. We modeled connections among the discourse codes as well, but we did not model connections among the errors, as the errors are unrelated and thus independent from each other. For ENA analyses of error management, only those residents who committed at least one error during the simulation were included ($n = 40$). Residents were grouped into low-performing ($n = 20$) and high-performing ($n = 20$) conditions by outcome score for analysis.

Results

Neither number of errors nor type of error predicts operative outcome

To assess whether the number of errors committed was predictive of outcome score, we performed a standard linear regression analysis. No correlation was found between the number of errors committed and outcome score ($R^2 = 0.03$).

To assess whether specific errors were predictive of outcome score, we performed a penalized linear regression analysis, which consists of a standard linear regression analysis with weighting on variables whose sample size falls below the threshold needed to prevent a faulty generalization based on separation variability at low sample sizes.[33] The only error which fell below this cutoff was *Drops Tool*. No single error was predictive of outcome score: *p*-values of > 0.05 were obtained for all six errors (Table 3).

Table 3: Linear Regression Table of Individual Errors against Outcome Score

Error	Estimate	SE	*t*-value	*p*-value
Insertion Without Visualization	2.47	1.62	1.53	0.14
No Incision Prior to Endoclose	1.67	1.62	1.03	0.31
Grabs Extra Suture	–1.79	2.18	–0.82	0.42
Two Sutures, Same Hole	0.05	1.91	0.03	0.98
Tacker Slips	2.39	1.60	1.49	0.15
Drops Tool	–4.05	2.71	–2.24	0.09

Surgeons' intraoperative speech in response to errors predicts outcome score

We performed an ENA analysis of residents' coded intraoperative discourse to determine how residents recognized and reacted to errors after they occurred. For this ENA model, we used a *moving stanza window*[34] of six utterances: codes that co-occurred within six utterances of one another were considered connected, while codes that occurred outside this window were not considered connected. Participants were grouped into low-performing ($n = 20$) and high-performing ($n = 20$) conditions by outcome score.

Figure 1 (right) shows the low- and high-performing residents (white and gray points, respectively) along with their mean locations in the network space, which are represented by white and gray squares. The associated with each mean indicate the 95% confidence interval on each dimension. This analysis showed a significant difference on the second (y) dimension in how low- and high-performing residents integrated skills and decision-making in response to errors ($\overline{y}_{LP} = 0.13$, $\overline{y}_{HP} = -0.08$; $t_y = 2.46$; $p_y = 0.02$; Cohen's $d_y = 0.83$).

To understand the differences between low- and high-performing residents, we constructed a difference graph (Figure 1, left). The difference graph subtracts the edge weights of the mean networks of the low- and high-performing residents, visualizing the differences in weights; connections represented by dashed lines were stronger among the low-performing residents, and connections represented by solid lines were stronger among the high-performing residents. As the difference graph shows, high-performing residents made proportionally more connections to *Operative Planning* after committing cognitive or visuospatial errors, and they made a stronger connection between *Identifying Errors* and *Operative Planning*. Low-performing residents were more likely to make operative plans only in response to *Motor Errors*, the simplest of the three types of error to identify and correct, and they were more likely to respond to errors with frustration. These differences indicate different levels of ability to integrate elements of error recognition and management in response to errors. Low-performing residents were less

able to identify and develop a plan to respond to errors than high-performing residents, particularly when the errors were visuospatial or cognitive.

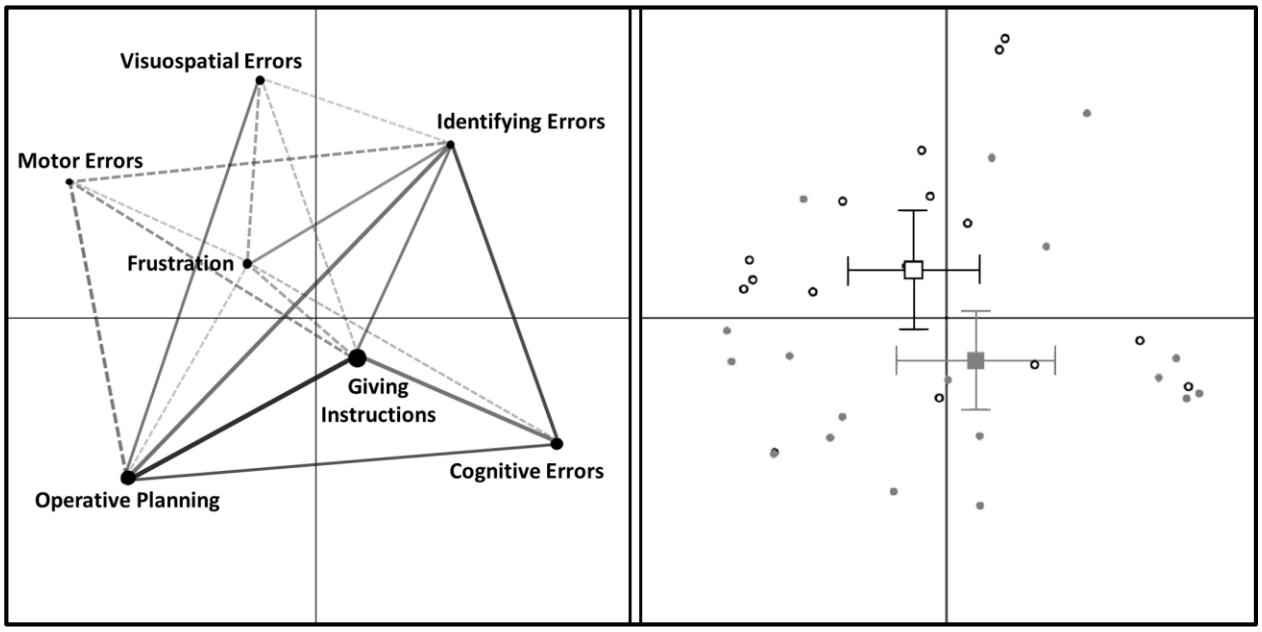

Figure 1: Comparison of error management between low- and high-performing residents using epistemic network analysis. The graph on the left shows a subtraction of the mean network of high-performing residents from the mean network of low-performing residents, which indicates the connections that were stronger among the low-performing residents (dashed lines) and those that were stronger among the high-performing residents (solid lines). Thicker lines indicate larger differences between the two groups. The graph on the right shows low-performing (white points, $n = 20$) and high-performing (gray points, $n = 20$) residents, with the corresponding means (squares) and 95% confidence intervals (bars). Each point is the centroid of one resident's network. The locations of the means provide the basis for measuring statistical differences between both groups of residents and are determined by the differences in network connection strengths. The network graph (left) enables interpretation of the statistically significant difference between the means of the two groups.

Discussion

This study examined a novel technique for modeling how general surgery residents from seven different institutions integrate elements of error recognition and management during an operative procedure. Residents who performed well on the final steps of a simulated LVH repair exhibited the same frequency and types of error as low-performing residents. However, high-performing residents were significantly more likely to manage complex errors by identifying the problem and verbalizing an operative plan to correct it. In other words, the quality of the hernia repair was significantly affected by how well residents' managed their errors.

Recent studies show that training residents not simply to avoid errors but to manage them can have positive effects on performance and on retention and transfer of skills.[35–38] In addition, research suggests that assessing performance by classifying the frequency of errors does not adequately capture important operative abilities, including error recognition, framing of adverse events, contingency planning, and error recovery, all of which are critical for operative independence.[39] Despite the development of numerous error classification checklists, assessment of error management remains largely

based on subjective observation and does not account for the ways in which surgeons integrate error management behaviors.

These results indicate that ENA can use error checklist data (procedure-specific) and surgeons' natural language discourse (procedure-agnostic) to model the integration of psychomotor, procedural, and cognitive elements of intraoperative performance as outlined in the framework proposed by Madani and colleagues.[3] That is, ENA can take qualitative, *in situ* data from an authentic operative context and construct a comprehensive, quantitative model of operative performance. Importantly, such models can be used both to provide formative feedback and to assess aptitude; most extant measures of operative ability are not suitable for both formative and summative assessment.[11–13]

In addition, ENA can provide actionable information about surgeons' strengths and weaknesses. Existing measures of surgical performance are based on procedure time and assessment of technical skills, which provide a limited picture of the complex abilities needed for operative independence.[14] Because ENA produces a network for every individual, as well as mean networks for selected groups, such models can be used to provide both individual feedback and population-level summaries of performance. At the individual level, ENA models indicate which elements of intraoperative performance a surgeon integrates well or poorly, providing targets for further training or practice. At the population level, such models can guide the development or improvement of curricula to address common shortcomings in the integration of skills, knowledge, and decision-making.

This study also has several limitations. While we provide evidence of model validity, further validation studies should be conducted. As we report above, the coding process is valid and reliable, the model identifies statistically significant differences between two groups with different outcomes, and the specific differences identified by the model correspond with hypothesized differences based on both a theoretical framework and qualitative analysis of the data. In future work, we will further validate this model by analyzing additional data. The approach described here should also be applied to additional operative procedures and contexts to assess its feasibility as a more general technique for modeling surgical performance.

The purpose of this preliminary study was to test a novel approach to modeling the integration of surgical skills, knowledge, and decision-making, particularly advanced cognitive skills related to error management. Ultimately, our goal is to produce models that provide comprehensive assessment of intraoperative performance. However, this study raises significant questions about which elements of operative competency to model and how to collect evidence that clearly documents those elements, and future research is needed to explore these questions in detail. For example, Madani and colleagues identify 21 distinct "behavioral themes" within the domain of advanced cognitive skills alone, and in this study, we examined only those related to "error/injury recognition, rescue, and recovery" (pp. 260-61).[3] There may be dozens of distinct behaviors that are relevant in a given context, each of which may have a different level of importance or may be modified in various ways depending on the procedure, on the details of the specific case, or on the surgeon's level of expertise. It may be necessary, then, to develop assessment models around specific clusters of behaviors—such as the error recognition and management model presented here—in order to make such assessments more tractable, and such models may require modification for different populations (e.g., junior residents, senior residents, &c.).

Even when behaviors are identified for inclusion in a model, there are important questions about how they should be defined and identified. For example, surgeons would generally agree that operative planning (both proactive and reactive) is a critical skill in virtually all operative contexts, and in the present study, we developed a reliable method for identifying reactive operative planning by applying regular expression matching to residents' intraoperative discourse. Our coding process, however, did not distinguish between appropriate and inappropriate plans—it identified only the behavior of planning, not whether the plan proposed was a good one. The extent to which the quality of a particular behavior needs to be modeled is an open question, and we will investigate this in future work. Another area where further research is needed involves the *order* of integration, as specific sequences of behaviors may be important in some contexts.

Similarly, determination of which errors to include in an analysis While the error checklist we applied in this study is based upon a validated set of criteria used to assess surgical competency on the LVH repair, the lack of impact errors had on outcome score may indicate insufficient breadth in the checklist. The size of the checklist is, in part, a byproduct of the simplified context in which participants performed the procedure. In other words, because residents were performing only the final steps of the LVH repair, the range of errors they could commit was somewhat limited. In addition, because this study is focused on error management, we omitted errors that were not committed by the residents in this study. In a longer procedure, a longer list of errors could be considered, and with a larger study population, a wider range of errors may occur.

The question of data collection is also non-trivial. For example, haptic data would likely provide the best evidence of surgeons' level of psychomotor skill. Metrics such as path length or smoothness of motion are excellent indicators of manual aptitude and confidence. However, collection of such data is not possible during live operations, unlike collection of audio-video or checklist data. Even if simulators are used, there is additional expense associated with the use of haptic sensors, and considerable data processing is required for meaningful interpretation. (Unlike audio data, which requires simple transcription to make it suitable for a wide range of analyses, haptic data must undergo extensive processing—e.g., sample-specific normalization, noise thresholding, and trajectory-supervised haptic rendering.) Thus, future work will need to explore the feasibility of collecting various kinds of performance data and evaluate the extent to which effective comprehensive models can be developed using only data that are relatively easy to collect; this is particularly necessary if for the development of assessment models that can be translated from research into practice.

That being said, the reader may wonder how difficult it is to conduct an ENA analysis of the kind described here. There are essentially three phases: (a) data collection and processing, (b) data coding, and (c) data analysis. Data collection involves audio and video recording and collection of the simulator skins. The audio is transcribed, a standard checklist is used to identify errors (live and/or from the video), and a standard rubric is used to score the quality of the repair using the skins. These processes are comparable to those used in many frameworks for assessing surgical performance. Data coding, as described above, requires some effort if new codes are to be generated, but once the codes are automated, then even very large datasets can be easily coded in a matter of seconds. Lastly, ENA analyses can be conducted using the free online ENA webkit (http://www.epistemicnetwork.org/). The ENA webkit supports analysis, visualization, and statistical hypothesis testing.

Conclusions

On a simulated LVH repair, high-performing residents exhibited the same frequency and types of error as low-performing residents. However, high-performing residents were significantly more likely to manage their errors effectively by integrating relevant skills, knowledge, and decision making. These results suggest that procedure-specific error checklist data and procedure-agnostic elements of intraoperative behavior can be used to model the integration of critical aspects of intraoperative performance. In addition, multi-modal ENA models provide actionable information about surgeons' strengths and weaknesses, which can inform the development of targeted educational interventions and improve the design of curricula to address common shortcomings.

Acknowledgments

This work was funded in part by the National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the Department of Defense (W81XWH-13-1-0080), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and

Graduate Education at the University of Wisconsin-Madison. A. R. Ruis was supported by the American College of Surgeons–University of Wisconsin Surgical Education Research Fellowship program. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

1. Makary MA, Daniel M. Medical error—The third leading cause of death in the US. *Br Med J*. 2016;353:.
2. Wiegmann DA. Analysis of Error Management Strategies during Cardiac Surgery: Theoretical and Practical Implications. Presented at the Thousand Oaks, CA.
3. Madani A, Vassiliou MC, Watanabe Y, *et al.* What are the principles that guide behaviors in the operating room? Creating a framework to define and measure performance. *Ann Surg*. 2017;265:255–267.
4. DiSessa AA. Knowledge in pieces. In: Forman G, Pufall P, eds. *Constructivism in the Computer Age*. Hillsdale, NJ: Erlbaum; 1988. p. 47–70.
5. Linn MC, Eylon B-S, Davis EA. The knowledge integration perspective on learning. In: Linn MC, Davis EA, Bell P, eds. *Internet Environments for Science Education*. Mahwah, NJ: Lawrence Erlbaum Associates; 2004. p. 29–46.
6. Shaffer DW. Models of situated action: Computer games and the problem of transfer. In: Steinkuehler C, Squire KD, Barab SA, eds. *Games, Learning, and Society: Learning and Meaning in the Digital Age*. Cambridge, UK: Cambridge University Press; 2012. p. 403–431.
7. Greenberg CC, Ghousseini HN, Quamme SRP, *et al.* Surgical coaching for individual performance improvement. *Ann Surg*. 2015;261:32–34.
8. Glarner CE, McDonald RJ, Smith AB, *et al.* Utilizing a novel tool for the comprehensive assessment of resident operative performance. *J Surg Educ*. 2013;70:813–820.
9. Martin JA, Regehr G, Reznick R, *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
10. Yule S, Paterson-Brown S. Surgeons' non-technical skills. *Surg Clin North Am*. 2012;92:37–50.
11. Anderson DD, Long S, Thomas GW, *et al.* Objective Structured Assessments of Technical Skills (OSATS) does not assess the quality of the surgical result effectively. *Clin Orthop*. 2016;474:874–881.
12. D'Angelo A-LD, Cohen ER, Kwan C, *et al.* Use of decision-based simulations to assess resident readiness for operative independence. *Am J Surg*. 2015;209:132–139.
13. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): A systematic review of validity evidence. *Adv Health Sci Educ*. 2015;20:1149–1175.
14. Schmitz CC, DaRosa D, Sullivan ME, *et al.* Development and verification of a taxonomy of assessment metrics for surgical technical skills. *Acad Med*. 2014;89:153–161.
15. Law Forsyth K, DiMarco SM, Jenewein CG, *et al.* Do errors and critical events relate to hernia repair outcomes? *Am J Surg*. 2017;213:652–655.
16. Shaffer DW, Hatfield DL, Svarovsky GN, *et al.* Epistemic network analysis: A prototype for 21st century assessment of learning. *Int J Learn Media*. 2009;1:1–21.
17. Shaffer DW, Collier W, Ruis AR. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *J Learn Anal*. 2016;3:9–45.
18. Shaffer DW, Ruis AR. Epistemic network analysis: A worked example of theory-based learning analytics. In: Lang C, Siemens G, Wise AF, Gasevic D, eds. *Handbook of Learning Analytics*. Society for Learning Analytics Research; 2017. p. 175–187.
19. Shaffer DW. *Quantitative ethnography*. Madison, WI: Cathcart Press; 2017.

20. Arastoopour G, Shaffer DW, Swiecki Z, *et al.* Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *Int J Eng Educ*. 2016;32:1492–1501.
21. Quardokus Fisher K, Hirshfield L, Siebert-Evenstone AL, *et al.* Network Analysis of Interactions between Students and an Instructor during Design Meetings.
22. Chesler NC, Ruis AR, Collier W, *et al.* A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *J Biomech Eng*. 2015;137:024701:1-8.
23. Bagley EA, Shaffer DW. Stop talking and type: Comparing virtual and face-to-face mentoring in an epistemic game. *J Comput Assist Learn*. 2015;26:369–393.
24. Nash P, Shaffer DW. Mentor modeling: The internalization of modeled professional thinking in an epistemic game. *J Comput Assist Learn*. 2011;27:173–189.
25. Hatfield DL. The right kind of telling: An analysis of feedback and learning in a journalism epistemic game. *Int J Gaming Comput-Mediat Simul*. 2015;7:1–23.
26. Wooldridge AR, Carayon P, Eagan BR, Shaffer DW. Quantifying the qualitative with epistemic network analysis: A human factors case study of task-allocation communication in a primary care team. *IIE Trans Healthc Syst Eng*. 2018:in press.
27. Sullivan SA, Warner-Hillard C, Eagan BR, *et al.* Using epistemic network analysis to identify targets for educational interventions in trauma team communication. *Surgery*. 2017:in press.
28. Pugh C, Plachta S, Auyang E, *et al.* Outcome measures for surgical simulators: Is the focus on technical skills the best approach? *Surgery*. 2010;147:646–654.
29. Peters JH, Fried GM, Swanstrom LL, *et al.* Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*. 2004;135:21–27.
30. Glaser BG, Strauss AL. *The discovery of grounded theory: Strategies for qualitative research*. Aldine de Gruyter; 1967.
31. Eagan BR, Rogers B, Pozen R, *et al. rhoR: Rho for inter rater reliability*. 2016.
32. Eagan BR, Rogers B, Serlin R, *et al.* Can We Rely on Reliability? Testing the Assumptions of Inter-Rater Reliability. Presented at the International Conference on Computer-Supported Collaborative Learning, Philadelphia, PA.
33. Kuhn M, Johnson K. *Applied predictive modeling*. Springer; 2013.
34. Siebert-Evenstone AL, Arastoopour G, Collier W, *et al.* In search of conversational grain size: Modeling semantic structure using moving stanza windows. In: Looi C-K, Polman J, Cress U, Reimann P, eds. *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016*, vol. I. 2016. p. 631–638.
35. Gardner AK, Abdelfattah K, Wiersch J, *et al.* Embracing errors in simulation-based training: The effect of error training on retention and transfer of central venous catheter skills. *J Surg Educ*. 2015;72:e158–e162.
36. Lorenzet SJ, Salas E, Tannenbaum SI. Benefiting from mistakes: The impact of guided errors on learning, performance, and self-efficacy. *Hum Resour Dev Q*. 2005;16:301–322.
37. Gully SM, Payne SC, Kiechel Koles KL, Whiteman J-AK. The impact of error training and individual differences on training outcomes: an attribute-treatment interaction perspective. *J Appl Psychol*. 2002;87:143–155.
38. Dror I. A novel approach to minimize error in the medical domain: Cognitive neuroscientific insights into training. *Med Teach*. 2011;33:34–38.
39. Law KE, Ray RD, D'Angelo A-LD, *et al.* Exploring senior residents' intraoperative error management strategies: A potential measure of performance improvement. *J Surg Educ*. 2016;73:e64–e70.