The American Journal of Surgery 220 (2020) 37–43

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Evaluating how residents talk and what it means for surgical performance in the simulation lab

Anne-Lise D. D'Angelo [a,*,1], Andrew R. Ruis [b], Wesley Collier [b], David Williamson Shaffer [b,c], Carla M. Pugh [a,2]

[a] University of Wisconsin-Madison, Department of Surgery, School of Medicine and Public Health, 600 Highland Ave, Madison, WI, 53792, USA
[b] University of Wisconsin-Madison, Wisconsin Center for Education Research, 489 Educational Sciences Bldg, 1025 West Johnson St, Madison, WI, 53706, USA
[c] University of Wisconsin-Madison, School of Education, Department of Educational Psychology, 1065 Educational Sciences Bldg, 1025 West Johnson St, Madison, WI, 53706, USA

A R T I C L E I N F O

Article history:
Received 10 May 2019
Received in revised form
22 January 2020
Accepted 2 February 2020

Keywords:
Simulation
Discourse analysis
Epistemic network analysis
Education
Assessment

A B S T R A C T

Background: This paper explores a method for assessing intraoperative performance by modeling how surgeons *integrate* skills and knowledge through discourse.
Methods: Senior residents (N = 11) were recorded while performing a simulated laparoscopic ventral hernia (LVH) repair. Audio transcripts were coded for five discourse elements related to knowledge, skills, and operative independence. Epistemic network analysis was used to model the ordered integration of the five discourse elements.
Results: Participants with poorer hernia repair outcomes had stronger connections between the discourse elements *operative planning* and *asking for information or advice* (Operative planning), while participants with better hernia repair outcomes had stronger connections between the discourse elements *giving assistant instructions* and *identifying errors* (Operative management): ($p = .006$; Cohen's $d = 2.79$).
Conclusion: Participants with better hernia repair outcomes engaged in more operative management communication during the simulated procedure. This ability to integrate multiple operative steps and verbally communicate them significantly correlated with better operative outcomes.

Introduction

The Halsted apprenticeship model of surgical education means that learning is situated in the context of surgical practice.[1] Clinical knowledge and skill is not gained in isolation, but rather through the care of patients.[2–4] Learning in this context requires interaction with other persons and technology and the integration of multiple pieces of knowledge and various skills in complex settings. In surgical practice, this interaction and integration must be translated into not only cognitive decisions, but also the physical execution of the task. However, current assessment measures in surgery do not evaluate how and to what extent surgeons achieve this integration in practice.

Surgical performance is often dichotomized into technical and non-technical skills. Technical skills involve proficiency or dexterity in the execution of a task or use of equipment.[5,6] Non-technical skills include elements such as teamwork, leadership, communication, situational awareness, and decision making.[7–9] Technical and non-technical skills are likely interrelated,[10–12] but many current assessment frameworks continue to make this distinction. Global technical skills are often evaluated with the Objective Structured Assessment of Technical Skills (OSATS),[13–15] while global non-technical skills are often assessed with the Non-technical Skills for Surgeons (NOTSS).[5,7–9] Notably, these assessment frameworks do not take into account how these elements are connected to each other. It is this connection between knowledge, skills, and actions that is critically important to the holistic assessment of surgical skills performance.[4,16]

* Corresponding author. Mayo Clinic – Rochester, Department of Colon and Rectal Surgery, 200 1st Street SW, Rochester, MN, 55902, USA.
E-mail address: dangelo.anne-lise@mayo.edu (A.-L.D. D'Angelo).
1 Present address: Mayo Clinic – Rochester, Department of Colon and Rectal Surgery, 200 1st, Street SW, Rochester, MN 55902.
2 Present address: Stanford University, Department of Surgery, 300 Pasteur Drive, Grant Building – Suite S067, Stanford, CA 94305.

https://doi.org/10.1016/j.amjsurg.2020.02.016
0002-9610/

A more complete way of evaluating surgical residents may involve assessing how trainees assimilate knowledge, skills, and interactions in the performance of a surgical task. Because surgeons' intraoperative talk provides insight into their understanding of a procedure, we hypothesized that qualitative discourse analysis can be used to assess the integration of technical and non-technical aspects of performance. Discourse analysis is the study of how people use language through written or spoken exchanges and can provide insight into thought processes and decision making.[17] To create models of resident discourse while performing a simulated procedure, we used epistemic network analysis (ENA), which is a technique for quantifying and visualizing the structure and strength of association among elements of complex task performance over time.[18–21] ENA has been successfully utilized to assess complex thinking and task performance in a wide variety of contexts,[22–27] including operative[28,29] and clinical[30] settings.

Materials and methods

Participants

Participants ($N = 20$) were senior residents (post-graduate year [PGY] 4 and 5) attending a two-day Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) advanced laparoscopic surgery course. This was a convenience sample based on attendance at the conference and participation in a simulation-based course. Study participants were predominately male (80%); more than half (60%) of the participants were PGY 4s, and the remaining participants (40%) were PGY 5s. As this study relies on discourse analysis, 9 of the participants were excluded because of poor audio and transcription quality or lack of an operative assistant during portions of the procedure. After this exclusion, 11 participants (90% male; 54% PGY 4) remained in the study population.

On day one of the SAGES course, residents completed a demographic survey indicating gender and post-graduate year. Next, they performed a simulated laparoscopic ventral hernia (LVH) repair. The operative goal was to use laparoscopy to repair the hernia defect with the use of mesh. The participants had 30 min to complete the simulated procedure. Surgical faculty acted as operative assistants during the procedure. These faculty members were instructed to take on the role of assistant, and the participant was instructed to act as the primary surgeon. There were ten simulation stations set up with individual faculty at each of the stations and two rounds of participants. All simulated procedures were video-recorded using both external and laparoscopic views. These video recordings served as the data for this study.

The University of Wisconsin–Madison Institutional Review Board (IRB) granted this study exempt status.

Laparoscopic ventral hernia simulator

The LVH simulator is a physical, box-style simulator designed to represent the abdominal cavity of a patient with a ventral hernia.[31] All necessary open and laparoscopic equipment, except cautery, was provided to complete the hernia repair using mesh. A 20×25 cm piece of mesh was provided.

Hernia completion score

Final product analysis was performed on the simulated LVH repairs to assess for completeness of the procedure and adequate coverage of the hernia. A blinded member of our research team used final product analysis to grade all participants' LVH simulator skins in four major areas: suture placement (maximum 6 points), tack placement (maximum 5 points), port placement (maximum 2 points), and mesh placement (maximum 4 points). The highest possible score is 17 points. A second blinded member of our research team also used final product analysis to grade a random sample (50%) of the same skins to assess inter-rater reliability. Inter-rater reliability, as measured by Cohen's kappa, was high ($\kappa = 0.83$).

Hernia completion scores were used to separate the participants ($n = 11$) into those with low hernia completion scores ($n = 4$) and high hernia completion scores ($n = 7$). A hernia completion score of 6 was selected as the cut off value with scores of 1–6 belonging to the low hernia completion scores group and scores of 7–17 belonging to the high hernia completion scores group. A hernia completion score of 6 was selected as the cut off because this was the median score for all original participants ($N = 20$) prior to exclusion criteria. The median was selected as this measure of central tendency is less likely to be skewed by outliers.

Discourse coding

Using a qualitative and iterative process,[21,28,32] we identified four discourse elements related to performance: ***operative planning***, ***identifying errors***, ***asking for information or advice***, and ***giving assistant instructions***. Additionally, we wanted to capture times when the assistant ***gives procedural advice***, as this may have influenced residents' performance. From these elements, we developed discourse codes, including code names, code descriptions, and code examples, for both the participants and assistants, which are detailed in Table 1. Discourse elements were identified and developed based on transcript review by three authors (AD, DWS, & CP). The elements were based on a qualitative assessment of the videos and expert knowledge of learning theory and surgical education.

Transcripts from the simulated procedure were segmented by turns-of-talk such that each turn of talk was represented as one line of data that was then coded. Speakers were identified as the participant, the assistant, or other. The participants' transcripts were coded by a single rater. Inter-rater reliability was established with a second blinded rater. The two raters independently coded a randomly selected subset of the data (10% of total turns of talk), and then agreement was assessed using Cohen's kappa. Inter-rater reliability measures for each of the codes are given in Table 1.

Epistemic network analysis (ENA)

After the data was coded, we used ENA software[33] to analyze how and to what extent residents integrated the coded discourse elements in their intraoperative speech. The theory and methodology of ENA has been explained in detail in prior publications.[19–21] In what follows, we explain only how ENA was applied in this study.

The ENA algorithm uses a moving window to construct a network for each line in the data—that is, for each turn of talk—showing how codes in a given line are connected to codes that occur within the recent temporal context.[34,35] Interpreting a given turn of talk requires identifying its appropriate relational context[36]: for example, the response to a question can only be fully understood in the context of the question itself. Researchers often define relational context by using a moving window: a number of prior turns of talk needed to understand a given turn of talk. The length of a given window is determined by the content and context of the discourse. Although each turn of talk may have a different number of referents—that is, a different number of prior turns to which it explicitly or implicitly refers—researchers use a moving window of fixed length to define the recent temporal context because prior turns of talk influence what participants say and do even when no reference is made. In this study, we used a window length of seven

Table 1
Code name, speaker type, description, examples and inter-rater reliability (κ) for discourse codes.

Code name	Speaker	Description	Examples	Inter-rater reliability (κ)
Operative planning	Participant	Participant states next procedural step, action or operative goal.	1) *Here's how I am going to try it. Let me see the um– Let me try and measure the sac size. Um. You want me to measure—* 2) *I'm just going to go fives from now on, since I'm just.* 3) *OK, good, I think I'm going to put a port here. So I'm just going to get the knife first. Can you just hold the camera a while? Just to triangulate.*	0.71
Identifying errors	Participant	Participant recognizes that a prior action or decision was incorrect or lead to a poor outcome	1) *This port site was too close to my hernia side. That's why.* 2) *Oopsy. It tore. See? Not going to hold.* 3) *The patient is going to have recurrence.*	0.71
Asking for information or advice	Participant	Participant seeks advice or information from the assistant relating to the procedure, equipment or simulator	1) *Yeah, I think its like 3 cm overlap here?* 2) *No, see, it slips right off. See, this isn't cinched down. Does it cinch down?* 3) *Yeah, what is your recommendation?*	0.75
Giving assistant instructions	Participant	Participant gives instructions to the assistant	1) *All right, look up one more time. That's fine, OK. Let's find the hernia pull back a little bit.* 2) *So push it up against that right there, and I'll bring that down. Let's come over here. You can grab that side. And then down.* 3) *I think it's open now. Show me the other stitch.*	0.85
Giving procedural advice	Assistant	Assistant provides advice or information relating to the procedure.	1) *You would not be able to get that into a five, so.* 2) *Your knot is on the back side. You want the knot up towards the abdomen.* 3) *OK, so you need to put in a second port, right?*	0.67

turns of talk. This means that for each turn of talk, the prior six turns of talk were considered within the window, and thus were considered to be connected. We selected a window length of seven based on our qualitative analysis, which indicated that this length reasonably captured the recent temporal context. Research also suggests that ENA models are relatively robust to the choice of window length.[35]

ENA produces a network for each line in the data based on the moving window. The network for each line indicates the unique connections between codes in that line and codes anywhere else in the window. These connections in these networks are binary, indicating only whether a connection appeared or did not appear in the window. To create a single network for each participant, the networks for each turn of talk by a given participant were summed (i.e., for each resident: n = 11). The result was 11 networks, one for each resident, that reflect the strength of association between each unique pair of codes.

To model not only the structure of connections among discourse elements but also the sequence with which those connections occurred, we preserved the order of connections in the ENA model. For example, in some instances, ***operative planning*** preceded ***asking for information or advice*** within a window, and in other instances, ***asking for information or advice*** preceded ***operative planning***. To account for this, in the network constructed for each line (i.e., for each turn of talk), codes that occurred in that line were labeled as *receiver*, and each unique code that appeared in the remainder of the window (i.e., in the prior six turns of talk) was labeled as *sender*. If discourse codes co-occurred within the same turn of talk (i.e., in the same line of data), then each code was labeled as both sender and receiver.

To control for the fact that different residents have different amounts of talk, and thus different numbers of coded lines in the data, we normalized the networks for all units of analysis by applying a sphere norm. Normalized networks were then subjected to a dimensional reduction (via singular value decomposition). The singular value decomposition was applied to the normalized data to produce orthogonal dimensions that maximize the variance explained by each dimension. This produces a network model that (a) enables comparison of residents' networks based on the *relative* strength of association among codes, and (b) facilitates

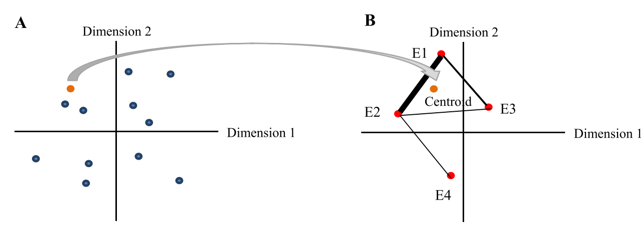

Fig. 1. A) Representation of multiple example participants' discourse network centroids plotted on ENA dimensions. The orange dot represents the discourse network centroid from Fig. 1B. B) Representation of a single example participant's discourse network with connections (black lines) between discourse elements (E1-E4, red circles) and centroid of the network graph (orange circle). The thickness of the connecting lines represents the frequency with which the discourse elements (E1-E4) are associated in stanzas. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

identification of the discourse connections that most differentiated low- and high-performers.

Networks were visualized using network graphs where the nodes correspond to the (ordered) codes—that is, there is a sender node and a receiver node for each of the five codes—and the graph edges, or connections between nodes, reflect the relative frequency of co-occurrence between two codes. The result is two coordinated representations for each unit of analysis: (1) an ENA score, which represents the location of a resident's network in the low-dimensional projected space (Fig. 1A), and (2) a weighted network graph (Fig. 1B). The positions of the network graph nodes are fixed, and those positions are determined by an optimization routine that minimizes the difference between each ENA score and the centroid of the network graph that corresponds with it. This is done so that the positions of the nodes in the network representations can be used to interpret the reduced dimensions, similar to principal components analysis.[37] For example, a participant with lower values on dimension 1 (x-axis) will have stronger connections between elements on the left side of the network graph, as represented by the example participant depicted in Fig. 1A and B. Because of this co-registration of network graphs and ENA scores, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the low-dimensional space and explain the positions of the ENA scores in the space. ENA can thus be used to compare units of analysis in terms of their ENA scores, individual networks, mean ENA scores, and mean networks, which average the edge weights across individual networks.

Significance testing

An independent-samples *t*-test was used to determine whether there was a significant difference between 1) the hernia completion scores of residents in the low-score group and those in the high-score group and 2) the operative discourse networks of residents with low hernia completion scores and those with high hernia completion scores.

Results

ENA model dimensions

Two dimensions of performance were analyzed based on the resultant ENA model. Dimension 1 (SVD 1) distinguishes between elements of **Independent performance** (*operative planning, giving assistant instructions,* and *identifying errors*) and **Assistance** (*asking for information or advice* and *assistant giving procedural advice*). Along dimension 2 (SVD 2), the elements of forward operative progression are further divided between *operative planning* (higher values on dimension 2) and *identifying errors* and *giving assistant instructions* (lower values on dimension 2). Thus, dimension 2 distinguishes between **Operative planning** (higher values) and **Operative management** (lower values).

Hernia repair outcomes

There was a significant difference between the mean hernia completion scores for the low hernia completion score group (n = 4; M = 3.8/17; SD = 0.96) and the high hernia completion score group (n = 7; M = 9.3/17; SD = 2.8): t(9) = −3.74, p = .005. This indicates that operative performance was statistically significantly different between the two groups. There was no significant differences in hernia completion scores between PGY 4s (n = 6; M = 7.33; SD = 4.5) and PGY 5s (n = 5; M = 7.20; SD = 2.59): p = .06.

Relationship between ENA model and operative performance

Fig. 2 shows the ENA scores of residents with low (yellow) and high (blue) hernia completion scores, along with the group means (large squares). There is a significant difference on dimension 2 (**Operative planning** vs **Operative management**) between residents with low hernia completion scores (M = 0.23, SD = 0.14) and residents with high hernia completion scores (M = −0.132; SD = 0.14): t(6.33) = 4.02, p = .006. Participants with low hernia repair scores had stronger connections between discourse elements *operative planning* and *asking for information or advice* (Operative planning), while participants with high hernia repair scores had stronger connections between discourse elements *giving assistant instructions* and *identifying errors* (Operative management). These results were based on an ordered model that took into account the temporal relationship between discourse elements. For example, strong connections between *operative planning* (sender) and *giving assistant instructions* (receiver); or between *identifying errors* (sender) and *asking assistant information or advice* (receiver) would create a network graph plotted in the direction of Operative management and Independent performance on the ENA dimensions (Fig. 2). The effect size, Cohen's d, is 2.79, which

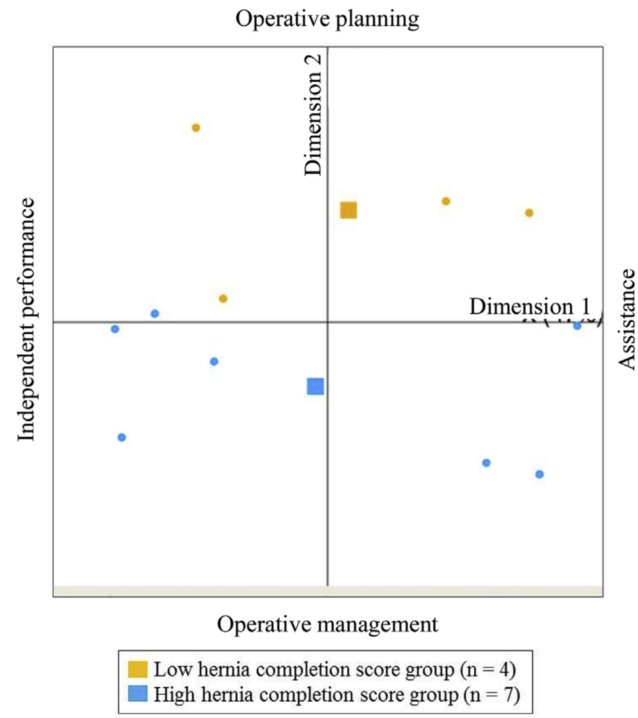

Fig. 2. ENA model showing the network locations of participants with low (yellow points) and high (blue points) hernia repair scores, along with the corresponding means (squares). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

represents a large difference between the two groups on dimension 2.

There is not a significant difference on dimension 1 (**Independent performance** vs **Assistance**) between residents with low hernia completion scores (M = 0.043; SD = 0.34) and residents with high hernia completion scores (M = −0.025; SD = 0.43), t (7.78) = −0.29, p = .78.

To demonstrate the relationship between the participant's discourse and the resultant ENA model, Fig. 3 shows all the individual participants plotted on the two ENA performance dimensions (Fig. 3A) with one participant's unique ENA network model (Fig. 3B) displayed along with a portion of their discourse along with how the discourse was segmented and the corresponding discourse codes (Fig. 3C). From qualitative analysis, this participant was noted to be proficient in ***operative planning***, as well as, ***identifying errors***, appropriately ***asking for information or advice*** from the assistant, and ***giving assistant instructions***. Importantly, they were able to integrate all of these elements together. The individual network demonstrates strong connections between the nodes ***operative planning, giving assistant instructions*** and ***identifying errors***. More specifically, there are strong connections from ***operative planning (sender)*** to ***identifying errors (receiver)***, ***giving assistant instructions (receiver)***, and ***asking for informaiton or advice (receiver)***. This indicates that the participant was able to create operative plans and then identify problems with those plans, provide the assistant instructions, and ask for information regarding the procedure. The participant was able to balance the different elements of operative performance. Additionally, there are strong connections from ***identifying errors (sender)*** to ***operative planning (receiver)*** and ***asking for informaiton or advice (receiver)***. That is, the participant responded to identified errors by asking for advice and formulating an operative plan. This participant has a fairly balanced model with multiple

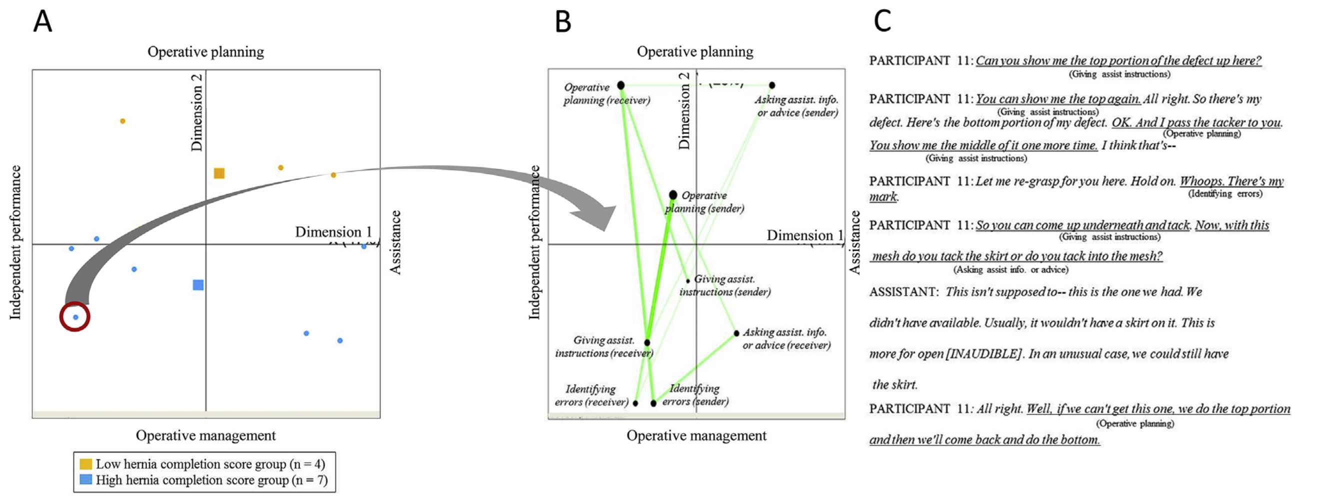

Fig. 3. A) ENA model with one participant selected (red circle). B) ENA network of that participant's discourse showing the structure of connections (lines) among discourse elements (nodes). The thickness and saturation of the connecting lines represents the relative frequency with which the discourse elements are associated. C) Excerpt from the transcript of that participant's discourse showing the coding of discourse elements. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

strong connections between multiple nodes, which is reflective of his ability to integrate different elements related to performing the procedure.

Discussion

Given that surgical training is situated in a complex environment with multiple interactions, surgical trainees should be assessed on how they assimilate knowledge, skills, and interactions in the performance of a surgical task. We hypothesized that mixed-methods discourse analysis can distinguish procedural outcomes during a simulated procedure. This study assessed the performance of chief residents during a simulated LVH repair using in-depth qualitative analysis, ENA modeling, and final product analysis. Our results demonstrated that ENA can distinguish performance during a simulated procedure by evaluating the connections between discourse elements found to be important for successful performance.

Using ENA, we created a network model that represented operative performance on two dimensions. Dimension 1 distinguished between elements of **Independent Performance** and **Assistance**, while dimension 2 distinguished between elements of **Operative planning** and **Operative management**. Using final product analysis (hernia repair scores), we separated participants into high and low performers. High performers had statistically stronger network connections in the area of operative management, while low performers engaged in more operative planning. This indicates that the elements of *identifying errors* and *giving assistant instructions* signify an ability that goes beyond simply deciding what to do next. Importantly it is this ability to integrate different aspects of **Operative management** that differentiates performance.

Interestingly, we did not see a significant difference between low and high performers on dimension 1 (**Independent performance** vs **Assistance**). One might think that participants with more connections indicative of **Independent performance** would have done better. However, the hernia completion score looks at the final product and does not take into account the assistance provided by other people. Those with higher hernia completion scores and high values on dimension 1 (**Assistance**) may have accomplished more in the procedure because of the additional assistance they received. This fact indicates that final product analysis should not be the only summative assessment measure. In this study, each participant had an operative assistant because this is not a procedure that can be done by a single surgeon alone. There was little overlap between the faculty members and participants so we were unable to analyze if one particular faculty member had a large influence on the participant's discourse or operative performance. By coding the interaction between participants and faculty through the discourse elements developed, we strived to capture this interaction which is indicated through dimension 1 (Independent performance vs Assistance). In future research, we will continue to explore additional discourse and behavioral factors that influenced the participant-faculty interaction and how this impacts operative performance. In simulation scenarios requiring an assistant or in group simulation scenarios it is important to take into account the interactions between the participant and the other person involved as this can be difficult to standardize, especially when the assistants are different people as in this study. This underscores the importance of not simply looking at the participants, but also the role of supporting persons when evaluating performance in a simulated environment.

The importance of the interaction between trainees and assistants has been highlighted with the development of the SIMPL (System for Improving and Measuring Procedural Learning) application for trainee assessment.[37] The SIMPL application was developed to track training progression in performance and provide trainees with timely and structured operative feedback from faculty. Aimed to provide multiple assessments of trainee performance throughout residency, a key factor in the SIMPL application is the four Zwisch stages of supervision starting with Show and Tell and progressing up to No Help.[38,39] This helps both trainees and faculty understand, where the trainee is in their ability to operate independently. At the final stage of No Help, the trainee can work with inexperienced first assistants, safely complete the case without faculty, recover most errors, and recognizes when to seek help/advice.[38,40] All of these aspects of independent performance require the discourse elements we identified in our ENA model - *operative planning*, *identifying errors*, *asking for information or advice* and *giving assistant instructions*. We found that trainees with higher hernia completion scores were able to engage in more operative management, specifically *identifying errors* and *giving assistant* *instructions*. To better understand which residents are ready for operative performance, we have to assess how trainees dictate the course of the operation through interactions with their operating partner. Reliance on case numbers will not achieve this goal.[41] Discourse analysis can play a vital role in elevating the level performance.

As with all research, there are limitations to this study. One limitation is the small number of participants. This limits our power to find statistically significant differences and generalize our results. The purpose of this study was not to generalize our model to the population of surgeons at large, or even to all residents completing a simulated LVH. Rather this analysis was performed to explore how we can model, and ultimately assess, operative performance in a way that accounts for the integration of technical and non-technical skills. This pilot study provides the basis for further investigation of this approach with larger and more diverse sample sizes. Another limitation to this study was the use of hand-coding to perform discourse analysis. Hand-coding is time intensive, and the use of this method limits the ability to perform it on a large scale. However, code development, inter-rater reliability assessment, and discourse coding in operative contexts can be performed with the use of automated coding algorithms.[28] The implementation of such a method would allow us to perform this analysis with more participants. With improvements in automated transcription and diarization of audio data, such an approach could soon be possible not only at scale but also in real time.

Given that operative performance does not occur in isolation, we must develop assessment methods that take into account the integration of procedural knowledge, technical skill, and communication. In this study, participants who performed well were able to evaluate the state of the procedure and utilize their knowledge to direct the assistant and identify errors. Operative management, or integrating multiple operative components in response to dynamic conditions, represents higher level operative knowledge, understanding, and decision making in contrast to operative planning alone.[37] Not only do these results inform the development of future assessment tools, they also direct where we can focus educational efforts. Future research and curriculum development efforts should concentrate on structured experiences that help to strategically advance residents from basic operative planning and step-by-step procedural knowledge to more complex operative management strategies.

Funding

This work was supported by a University of Wisconsin, Department of Surgery institutional grant; and the National Institutes of Health F32 grant [#F32EB017084-01].

References

1. Lave J, Wenger E. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press; 1991.
2. Brown JS, Collins A, Duguid P. Situated cognition and the culture of learning. *Educ Res*. 1989;18(1):32–42.
3. Ruis AR, Shaffer DW, Shirley DK, Safdar N. Teaching health care workers to adopt a systems perspective for improved control and prevention of healthcare-associated infections. *Am J Infect Contr*. 2016;44(11):1360–1364.
4. 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(2):255–267.
5. Flin R, Yule S, Paterson-Brown S, et al. Teaching surgeons about non-technical skills. *Surgeon*. 2007;5(2):86–89.
6. Satava RM, Gallagher AG, Pellegrini CA. Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. *J Am Coll Surg*. 2003;196(6):933–937.
7. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery*. 2006;139(2):140–149.
8. Yule S, Flin R, Maran N, et al. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World J Surg*. 2008;32(4):548–556.
9. Yule S, Paterson-Brown S. Surgeons' non-technical skills. *Surg Clin*. 2012;92(1):37–50.
10. Johnson AP, Aggarwal R. Assessment of non-technical skills. Why aren't we there yet? *BMJ Qual Saf*. 2019;28(8):606–608.
11. Mishra A, Catchpole K, Dale T, et al. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surg Endosc*. 2008;22(1):68–73.
12. Pfandler M, Stefan P, Mehren C, et al. *Technical and Non-technical Skills in Surgery: A Simulated Operating Room Environment Study*. Spine; 2019 (Epub ahead of print).
13. MacRae H, Regehr G, Leadbetter W, Reznick R. A comprehensive examination for senior surgical residents. *Am J Surg*. 2000;179(3):190–193.
14. Martin JA, Regehr R, Reznick H, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273–278.
15. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173(3):226–230.
16. Pugh CM. A holistic model of surgical expertise and competency. *Ann Surg*. 2017;265(2):268–269.
17. Gee JP. *An Introduction to Discourse Analysis: Theory and Method*. London, UK: Routledge; 1999.
18. 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.
19. 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.
20. 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; 2017:175–187.
21. Shaffer DW. *Quantitative Ethnography*. Madison, WI: Cathcart Press; 2017.
22. Zhang S, Liu Q, Cai Z. Exploring primary school teachers' technological pedagogical content knowledge (TPACK) in online collaborative discourse: an epistemic network analysis. *Br J Educ Technol*. 2019.
23. Swiecki Z, Ruis AR, Farrell C, Shaffer DW. Assessing individual contributions to collaborative problem solving: a network analysis approach. *Comput Hum Behav*. 2019. In press.
24. Csanadi A, Eagan B, Shaffer DW, et al. When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. *International Journal of Computer-Supported Collaborative Learning*. 2018;13(4):419–438.
25. Andrist S, Ruis AR, Shaffer DW. A network analytic approach to gaze coordination during a collaborative task. *Comput Hum Behav*. 2018;89:339–348.
26. 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(3B):1492–1501.
27. 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(2).
28. Ruis A, Rosser AA, Quandt-Walle C, et al. The hands and head of a surgeon: modeling operative competency with multimodel epistemic network analysis. *Am J Surg*. 2018;216(5):835–840.
29. Sullivan SA, Warner-Hillard C, Eagan B, et al. Using epistemic network analysis to identify targets for educational interventions in trauma team communication. *Surgery*. 2018;163(4):938–943.
30. Wooldridge AR, Carayon P, Shaffer DW, Eagan B. Quantifying the qualitative with epistemic network analysis: a human factors case study of task-allocation communication in a primary care team. *IISE Transactions on Healthcare Systems Engineering*. 2018;8(1):72–82.
31. 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(5):646–654.
32. Shaffer DW. Epistemic frames and islands of expertise: learning from infusion experiences. In: Kafai Y, Sandoval WA, Enyedy N, Nixon AS, Herrera F, eds. *Proceedings of the Sixth International Conference of the Learning Scienc. Santa Monica, CA*. 2004:473–480.
33. Marquart CL, Hinojosa C, Swiecki Z, et al. Epistemic network analysis [software]. http://app.epistemicnetwork.org; 2018.
34. Siebert-Evenstone A, Arastoopour Irgens G, Collier W, et al. In search of conversational grain size: modelling semantic structure using moving stanza windows. *Journal of Learning Analytics*. 2017;4(3):123–139.
35. Ruis AR, Siebert-Evenstone AL, Pozen R, et al. Finding common ground: a method for measuring recent temporal context in analyses of complex, collaborative thinking. In: LundNiccolai KG, Lavoué E, Hmelo-Silver C, Gwon G, Baker M, eds. *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings: 13th International Conference on Computer Supported Collaborative Learning*. CSCL; 2019.
36. Arvaja M, Salovaara H, Häkkinen P, Järvelä S. Combining individual and group-level perspectives for studying collaborative knowledge construction in context. *Learn InStruct*. 2007;17(4):448–459.
37. Arastoopour G, Swiecki Z, Chesler NC, Shaffer DW. Epistemic network analysis as a tool for engineering design assessment. In: *Paper Presented at the American Society for Engineering Education. Seattle, WA*. 2015.
38. Bonen JD, George BC, Williams RG, et al. The feasibility of real-time

intraoperative performance assessment with SIMPL (system for improving and measuring procedural learning): early experience from a multi-institutional trial. *J Surg Educ*. 2016;73(6):e118–e130.
39. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. *J Surg Educ*. 2013;70(1):24–30.
40. Williams RG, Swanson DB, Fryer JP, et al. How many observations are needed to assess a surgical trainee's state of operative competency? *Ann Surg*. 2019;269(2):377–382.
41. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ*. 2014;71(6):e90–96.