Education Management Platform Enables Delivery and Comparison of Multiple Evaluation Types

Ruchi M. Thanawala, MD,[*,†] Jonathan L. Jesneck, PhD,[*,†] and Neal E. Seymour, MD[*,†]

[*]University of Massachusetts Medical School-Baystate, Springfield, Massachusetts; and [†]University of Iowa Health Care, Carver College of Medicine, Iowa City Iowa

OBJECTIVE: The purpose of this study was to determine whether an automated platform for evaluation selection and delivery would increase participation from surgical teaching faculty in submitting resident operative performance evaluations.

DESIGN: We built a HIPAA-compliant, web-based platform to track resident operative assignments and to link embedded evaluation instruments to procedure type. The platform matched appropriate evaluations to surgeons' scheduled procedures, and delivered multiple evaluation types, including Ottawa Surgical Competency Operating Room Evaluation (O-Score) evaluations and Operative Performance Rating System (OPRS) evaluations. Prompts to complete evaluations were made through a system of automatic electronic notifications. We compared the time spent in the platform to achieve evaluation completion. As a metric for the platform's effect on faculty participation, we considered a task that would typically be infeasible without workflow optimization: the evaluator could choose to complete multiple, complementary evaluations for the same resident in the same case. For those cases with multiple evaluations, correlation was analyzed by Spearman rank test. Evaluation data were compared between PGY levels using repeated measures ANOVA.

SETTING: The study took place at 4 general surgery residency programs: The University of Massachusetts Medical School-Baystate, the University of Connecticut School or Medicine, the University of Iowa Carver College of Medicine, and Maimonides Medical Center.

PARTICIPANTS: From March 2017 to February 2019, the study included 70 surgical teaching faculty and 101 general surgery residents.

RESULTS: Faculty completed 1230 O-Score evaluations and 106 OPRS evaluations. Evaluations were completed quickly, with a median time of 36 ± 18 seconds for O-Score evaluations, and 53 ± 51 seconds for OPRS evaluations. 89% of O-Score and 55% of OPRS evaluations were completed without optional comments within one minute, and 99% of O-Score and 82% of OPRS evaluations were completed within 2 minutes. For cases eligible for both evaluation types, attendings completed both evaluations on 74 of 221 (33%) of these cases. These paired evaluations strongly correlated on resident performance (Spearman coefficient = 0.84, $p < 0.00001$). Both evaluation types stratified operative skill level by program year ($p < 0.00001$).

CONCLUSIONS: Evaluation initiatives can be hampered by the challenge of making multiple surgical evaluation instruments available when needed for appropriate clinical situations, including specific case types. As a test of the optimized evaluation workflow, and to lay the groundwork for future data-driven design of evaluations, we tested the impact of simultaneously delivering 2 evaluation instruments via a secure web-based education platform. We measured the evaluation completion rates of faculty surgeon evaluators when rating resident operative performance, and how effectively the results of evaluation could be analyzed and compared, taking advantage of a highly integrated management of the evaluative information. (J Surg Ed 76:e209–e216. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Resident operative evaluations, Multiple evaluations, O-Score, OPRS, Surgical data management

Funding source: None

Correspondence: Inquiries to: Neal Seymour, MD, University of Massachusetts Medical School-Baystate, 759 Chestnut Street, Springfield, MA 01199; e-mail: Neal.Seymour@baystatehealth.org

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2019.08.017

COMPETENCIES: Professionalism, Practice-Based Learning and Improvement, Systems-Based Practice

INTRODUCTION

Evaluation of performance is an essential responsibility of the teaching faculty members of any surgical residency. There is agreement that the absence of effective feedback is an impediment to high-quality medical training,[1] and that frequent evaluations are required for effective resident assessment.[2-5] Although the use of evaluation as a feedback tool is vitally important in surgical training, implementing a successful evaluation program requires overcoming many daunting challenges. Faculty members frequently juggle conflicting responsibilities, documentation burnout, and ever-increasing workload, and time pressures. Those pressures can push evaluation activities to the backburner, cause poor compliance with thoughtful evaluation completion, and result in a low evaluation completion rate. Some faculty members may not dedicate time for evaluations, without an obvious feedback loop to demonstrate a clear benefit on resident learning. When the evaluations are completed after a long delay (especially after the resident has moved on to another rotation), the potential for effective learning is further lessened, which might in turn intensify an evaluator's perception of lack of value.

In addition to faculty participation challenges, a significant logistics hurdle must be overcome to ensure that the appropriate evaluations are delivered to the right people at the right time. Manual delivery strategies, such as with paper-based evaluations or asking educators to remember to open an app to do an evaluation, suffer from being too separated from clinical workflow ("out of sight, out of mind"). This is especially pertinent for formative evaluations of individual resident case experiences, which must compete with the rush of necessary clinical and documentation responsibilities during the frequently busy and pressured postoperative period. As a partial solution, some evaluation strategies place the burden of initiating and requesting the evaluation upon the resident.[6] This strategy presents 2 potentially confounding selection biases. The first is a conflict of interest, since after a poor performance, some residents might elect not to request an evaluation in order to avoid having their performance documented. The second selection bias stems from the overall demands of surgery residency. Residents who struggle with residency demands and workflow habits might not spare the time or mental bandwidth for supposedly "non-essential" tasks, such as participation in evaluation. These confounding factors present an incomplete or inaccurate view of resident performance.

The program director is also confronted with the significant challenge of choosing from a large number of evaluation types and methods while advocating for timely completion of evaluations and control of evaluation quality. Although the Accreditation Council for Graduate Medical Education (ACGME) explicitly defines this responsibility in section V of the Common Program Requirements,[7] it leaves significant latitude for training programs to choose their evaluation methods. From a purely practical standpoint, the most useful system of evaluation is one that the evaluators will be most apt to use.[8] Ideally, it should also deliver an assessment that is appropriate to the case and the learner and include sufficient detail to enable a directed course of action for improvement without being excessively long or complex. Evaluation types cover a spectrum of characteristics from generic and minimalistic, which might facilitate completion,[9] to more richly detailed, which places a larger completion burden on educators and might apply to more specific settings.[10] Examples might include procedure-specific operative evaluation types or assessments of nontechnical skills such as history-taking. In theory, educators could provide extensive performance information by choosing appropriate evaluations from a large inventory of established evaluation types. However, it is impractical to expect busy faculty members to explore and choose from a large inventory on a frequent basis.

In considering the complexity inherent in parsing varied evaluation types, matching them to individual cases and residents, and making them available when most needed, we applied computational methods to lay the groundwork for automated selection and delivery of the evaluations. This study is an early step toward our long-term goal of using artificial intelligence to design tailored evaluations, based on the procedural experience and learning needs of individual residents.

MATERIAL AND METHODS

We built a secure, HIPAA-compliant, web-based platform for resident education management (Fig. 1).[11,12] The platform facilitated, and tracked several aspects of resident education and performance, including case assignments, case logging, case outcomes, reading of targeted educational materials, and operative performance evaluations. The platform synced with operating room (OR) schedules and resident service rotation schedules to enable live case assignments and automatic matching of case details with evaluations. Based on the case procedure details and case staff, the platform identified relevant evaluations from a bank of available evaluations, including the Ottawa O-Score instrument rating of

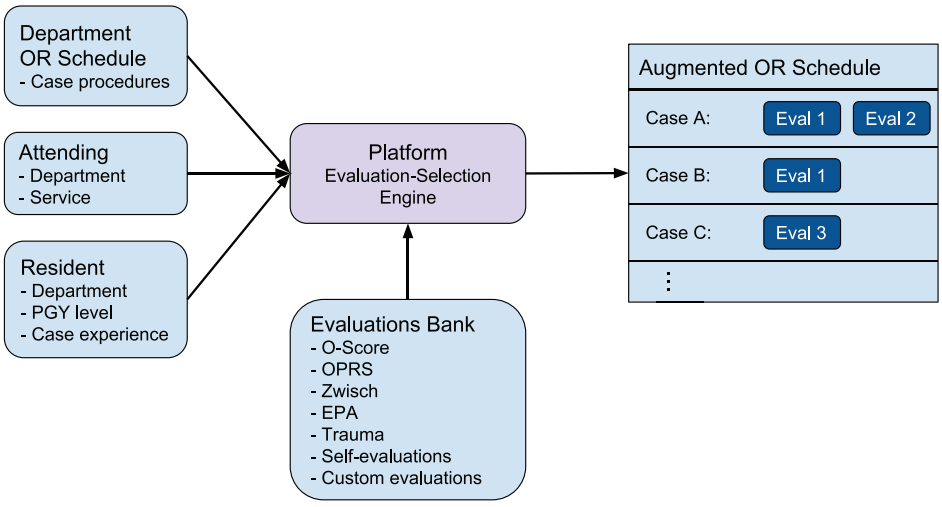

FIGURE 1. The platform integrates data from the OR schedule and assigned case staff, along with a data bank of available evaluations, to find appropriate evaluation and match them to each teaching case.

operative autonomy,[13] Operative Performance Rating Systems (OPRS) evaluations,[14] Entrustable Professional Activities (EPA) evaluations, and resident self-evaluations. All evaluations were automatically paired with appropriate teaching cases, and layered onto the operative schedule, where faculty and residents could easily find them and work them into their daily workflow. Faculty could choose whether to fill out one or more appropriate evaluations for each teaching case. For any teaching cases that still needed evaluations at the end of each day, the platform automatically sent brief reminder emails to the attendings to complete the evaluations, and upon completion, it immediately pushed the evaluation results to the residents. Evaluation results were streamed into resident performance dashboards for residents, faculty, and program directors. The dashboards tracked resident learning with case experience, operative performance, and progress toward ACGME requirements. The platform has been deployed multi-institutionally and across several departments.

For an initial test of the evaluation data quality, we measured the ability of the operative scores to stratify the residents by their program year (PGY) levels. Then, our principal measure of the usability of the platform's evaluation system was the time faculty spent to complete the evaluations. Each evaluation was structured as a short set of Likert-scale questions, followed by optional comments. We split the evaluation responses into 2 sets, those with and those without comments, and on each set, we measured the distribution of completion time using a Student's *t* test with unequal variance, and linear models.

Delivering multiple appropriate evaluations together for the same cases afforded a unique opportunity to study resident performance across operative evaluation types. We identified cases where faculty completed both the procedure-agnostic O-Score and procedure-specific OPRS evaluations on the same resident. For these matching evaluations, we measured the Spearman rank-order correlation of the resident overall operative performance. We also investigated whether faculty completed both evaluations together in one sitting, or at separate times. We measured the evaluation lag as the number of days between the teaching case and then submission of the corresponding evaluation. Finally, we explored correlations between pairs of questions across the evaluations.

RESULTS

The study included 70 teaching faculty members and 101 residents from 4 general surgery residency programs, starting with the University of Massachusetts Medical School-Baystate and then later expanding to include the University of Connecticut School of Medicine, the University of Iowa Carver College of Medicine, and Maimonides Medical Center. With 4546 teaching cases from March 2017 to February 2019, 33 attendings completed 1230 O-Score evaluations, 106 OPRS evaluations, and 14 EPA evaluations for 67 residents. Residents also completed 629 self-evaluations. Both the O-Score and OPRS evaluations stratified the residents by their program year, although in the OPRS evaluations, the attendings assigned slightly higher autonomy scores to the residents than expected (Fig. 2).

Evaluations were completed quickly, with the completion time depending mostly on the level of detail that the attending chose to include in the optional comments. For evaluations without comments, the median completion times were 36 ± 18 seconds for O-Score

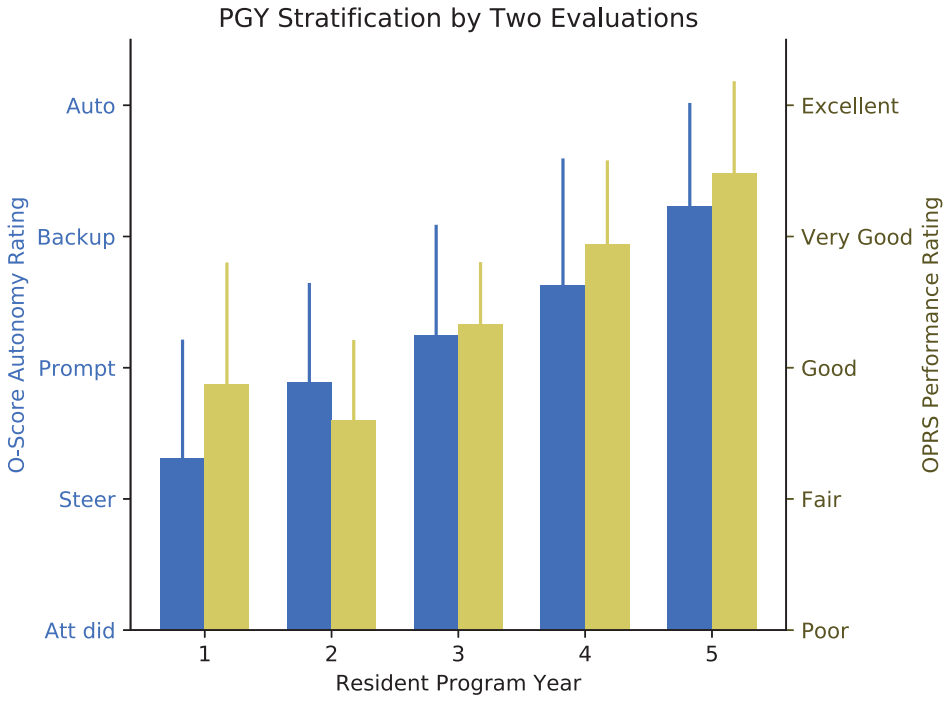

FIGURE 2. The 2 evaluations stratified the residents across program year levels ($p < 0.0001$). A larger average ORPS performance score for PGY 1 residents could have resulted from less complex cases appropriate for beginning surgery residents.

evaluations, and 53 ± 51 seconds for the OPRS evaluations. For evaluations with comments, the times increased to 1.79 ± 1.12 minutes for O-Score and 1.87 ± 1.09 minutes for OPRS (t-test with unequal variance, $p < 0.00001$) (Fig. 3). The overall evaluation completion time varied approximately linearly with comment length ($r = 0.85$, $p < 0.00001$ for O-Score, and $r = 0.54$, $p = 0.001$ for OPRS) (Fig. 4).

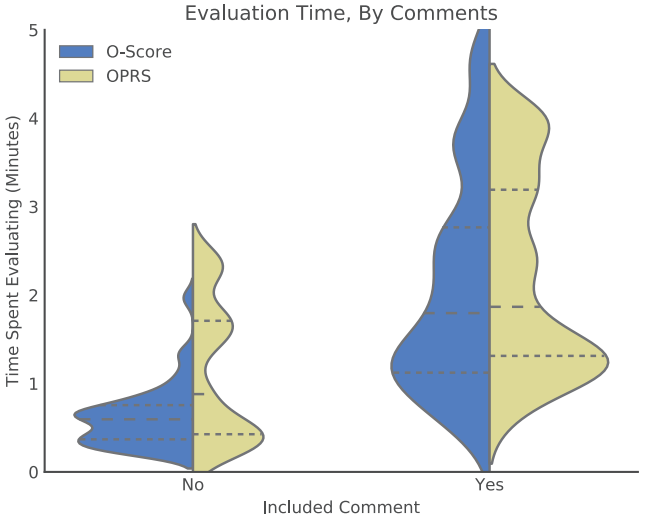

FIGURE 3. Faculty completed the evaluations quickly, especially when they opted not to include the optional comments ($p < 0.00001$).

There were 221 teaching cases that were eligible for both the O-Score and OPRS evaluations for the same resident, allowing for direct analysis of the timing and scoring across the paired evaluations (Table 1). Faculty completed both evaluation types for 74 of 221 (33%) eligible cases and completed at least one evaluation type for 141 (64%) cases, indicating a strong preference for completing both evaluation types (Fisher exact test, p-value < 0.00001). Faculty almost always completed both evaluations in one session, within a few days of the case (robust linear regression, $r = 0.97$, $p < 0.0001$) (Fig. 5) and within 1 minute ± 38 seconds of each other (Fig. 6). The paired evaluations showed a high correlation for resident overall operative performance (Spearman's rho = 0.84, $p < 0.00001$) (Fig. 7). We measured the correlation across all pairs of questions across evaluations (Fig. 8). The pairwise correlations were consistently high (rho > 0.7) with the exception of knot tying, which showed very little correlation across the other skills.

CONCLUSIONS

In light of the systemic challenges in flexibly providing a range of evaluation types, we framed the evaluation-feedback problem as a Surgery Department-wide information flow network, with resident performance information flowing via evaluations from surgical educators to program directors and residents, who communicate

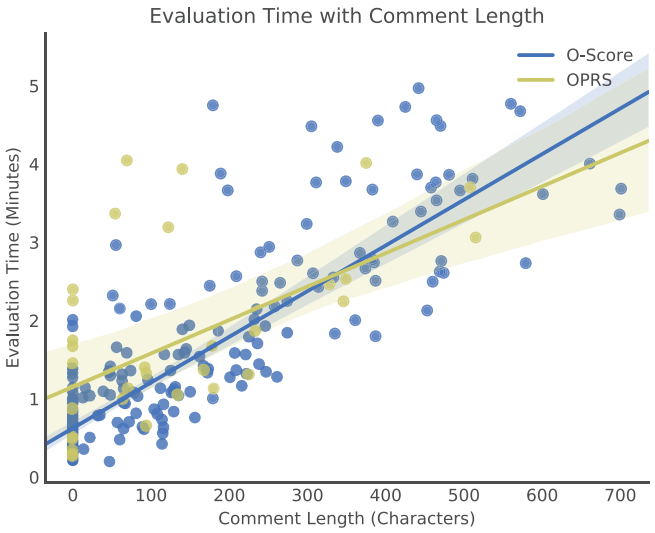

FIGURE 4. Most of the evaluation time was due to writing comments ($p < 0.0001$).

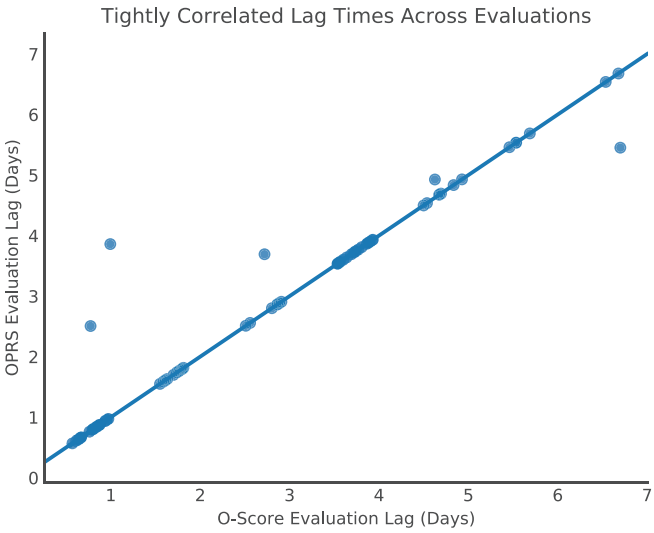

FIGURE 5. Faculty almost always completed both evaluations together, within a few days of performing the case with the resident ($p < 0.0001$).

TABLE 1. Dual-Evaluation Completion Rates. Of the 221 teaching cases that were eligible for both the O-Score and OPRS evaluations, faculty completed both evaluations for 74 cases (33%) and completed at Least one evaluation for 141 (64%) cases, indicating a strong preference for completing both evaluation types (p-value < 0.00001)

Cases	With O-Score	Without O-Score	Total
With OPRS	74	29	103
Without OPRS	38	80	118
Total	112	109	221

information back to educators by their performance in subsequent cases. With the challenges modeled as flow bottlenecks, we sought to facilitate an evaluation program by using process automation, workflow engineering, and artificial intelligence. Increasing faculty member efforts to complete evaluations would require an intelligent system to understand the operative schedule in order to identify evaluation opportunities, automatically select evaluations appropriate to the cases and the residents, and automatically deliver the evaluations to the teaching faculty members. To encourage participation, the evaluator would need a simple, intuitive front-end experience with a highly-optimized workflow. A back-end evaluation-processing system would process the entered information, track resident performance trends, and measure progress, and present evaluation results via intuitive plots in a dashboard, to make it more accessible and actionable to both the learner and to the education leadership infrastructure. We created such an automated evaluation selection, delivery, and tracking system, and then sought to measure its impact on evaluation completion. To measure the completion rates in this study, we considered cases where the evaluator could complete multiple types of evaluations for the same case, a task that would be excessively time-consuming without effective workflow optimization.

Addressing the described bottleneck problems, the platform enabled flexibility in the evaluation of resident operative performance. By integrating the data from the department OR schedules, faculty staff profiles, resident profiles, and multiple types of evaluations, the platform automatically identified teaching cases and matched them with the appropriate evaluations. By reducing the friction in the evaluation selection and delivery process, it was easier for time-pressed faculty members to participate and complete their evaluations, even multiple evaluations per case. The evaluation process was improved for all 3 relevant groups: (1) Faculty members see appropriate evaluations in their personal operative schedules and get automated reminder emails, (2) Residents get much more timely feedback on their performance, and they do not have to initiate set-up work to create the evaluations or send them to their attendings, and (3) Program directors experience much higher compliance rates from their teaching faculty and see their residents' performance trends in a real-time dashboard.

The platform facilitated faculty members' use of evaluations of individual residents' operative case performance. Before the platform's availability, the default practice had been to participate only in summative end-of-rotation evaluations, with the exception of OPRS evaluations required for graduating chief residents. The ease of access to the operative case-based evaluations via the platform facilitated a

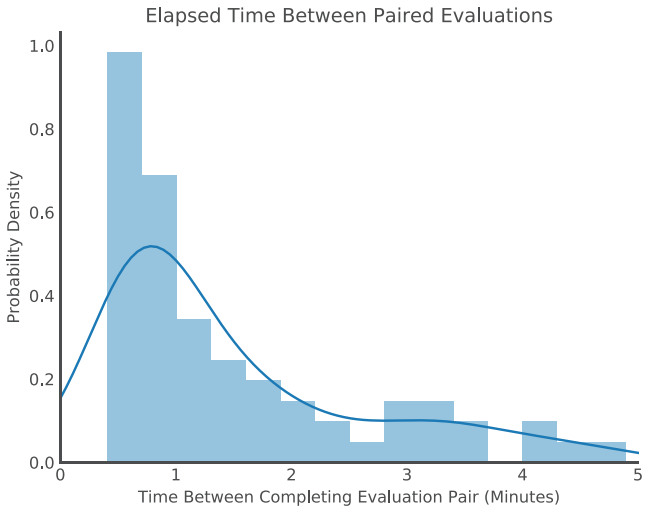

FIGURE 6. For the paired O-Score and OPRS evaluations, faculty completed evaluations in rapid sequence, within 1 minute ± 38 seconds.

workflow change and new habit formation for approximately half of the teaching faculty members. The case-based evaluation rate increased from near 0% to approximately 27% of all teaching cases. Despite the variable wait times across the group for each faculty member to submit their first evaluation, most participating faculty demonstrated a habit change, continuing to evaluate on subsequent teaching cases. We found a consistent impact on evaluation completion across all 4 surgery residency programs.

The proactive delivery and subminute completion times of the evaluations are likely explanations for their sustained use. The 12-question O-Score evaluation was straightforward enough that many faculty members were willing to add it to their postoperative workflow. Longer completion times were expected for the procedure-specific OPRS (approximately 20 questions depending on the procedure type). However, median OPRS evaluation time was in fact still under 1 minute. Because the EPA evaluations were deployed at only one institution (UConn) and relatively late in the study, we did not collect a sufficient number of evaluations for meaningful analysis. Overall, however, we believe that Likert-scale evaluations, conveniently presented to faculty evaluators, were short and quick enough to enable them to be folded into a daily workflow, and that the evaluation comments allowed for feedback and guidance to the residents. The corollary goal was to enable a positive feedback-and-learning cycle, where faculty would participate more, feeling that their evaluating time was valuable, as their feedback was delivered to the residents in real time soon after each case. It is hoped that earlier feedback facilitates resident performance improvements. Measuring its impact will be a focus of an upcoming study.

Confirmation of evaluator and resident satisfaction with platform use and workflow changes, as well as analysis of evaluation results quality will require an additional study. However, since we felt that it was typically rare for faculty members to fill out multiple evaluation types on the same case, cases that were eligible for multiple evaluations provided a testing ground that was particularly sensitive to examining workflow optimization. By offering multiple evaluation types where appropriate within the operative schedule, the platform enabled faculty to complete both O-Score and OPRS evaluations on

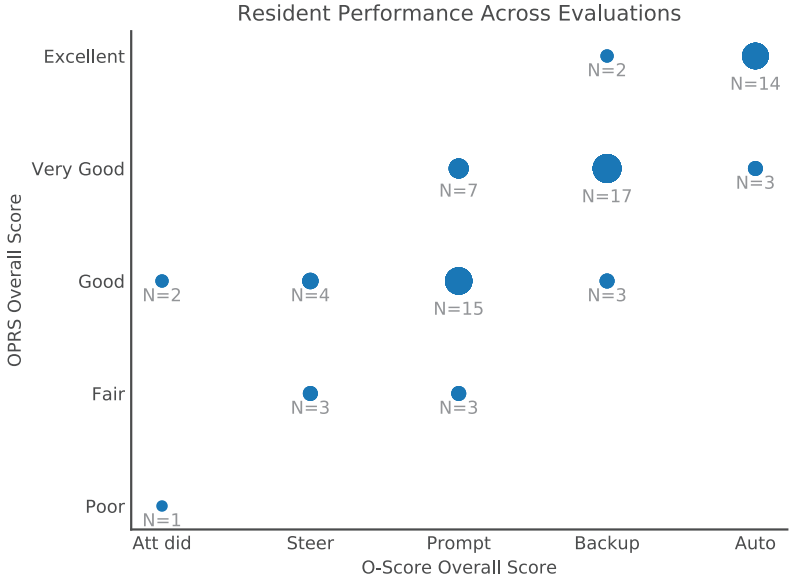

FIGURE 7. Paired O-Score and OPRS evaluations showed high correlation (rho = 0.84, $p < 0.00001$) for resident overall operative performance. The size of the dots indicates the number of matching evaluations at each score level.

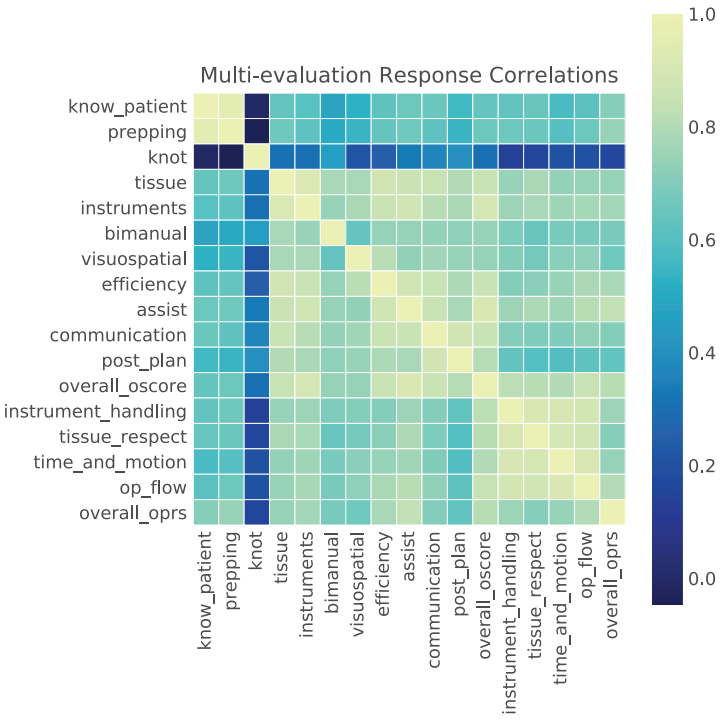

FIGURE 8. Comparing questions across multiple matched evaluations enables a detailed view of the response patterns. In this subset, most questions demonstrated moderate correlation, with the exception of knot tying. Perhaps because knot tying is an early-level mechanical skill, it did not correlate with broader skills that require more experience and background knowledge.

a third of the eligible cases. The paired evaluations demonstrated generally high correlations across their questions, indicating a well-balanced skills progression as residents gained operative experience. However, the notable outlier was knot tying, which showed no correlation to the other skills. Perhaps knot tying is a mechanical skill that is taught early in surgery residency and can be practiced in isolation, before the resident has the experience or background knowledge needed for higher-level skills, such as planning and decision-making in the OR, and general efficiency with time and motion during procedural steps. By comparing questions from several evaluation sources, it becomes possible to find an optimal set of predictive questions that minimize the faculty burden, and therefore maximize faculty participation, and maximize actionable utility to the residents. Multievaluation data collected in a large scale can possibly reorient and accelerate the evaluation design process. Rather than carrying out a prolonged study to validate a fixed evaluation, a platform that continuously tracks faculty participation and resident performance improvement could enable a "rolling" strategy for prioritizing and selecting informative and actionable questions from several sources and packaging them into optimal, short evaluations delivered to the right faculty at the right time in their residents' educational journeys.

The study had several limitations. Some faculty members opted out of submitting digital evaluations, either to avoid the legal risk of documenting resident operative mistakes in the event of a bad surgical outcome, or because they preferred to give their feedback to residents by other means. The O-Score and OPRS evaluations were not designed for resident teaching assistant (TA) cases, and there were reports of confusion about how to evaluate a chief resident guiding a junior resident through a case. Additionally, it was difficult for faculty to judge full operative autonomy in situations where the resident did only selected portions of the case. Because the platform integrated with the OR schedule, there was an inherent bias toward scheduled cases rather than add-on cases, which had to be manually entered before an evaluation was available. This scenario resulted in some of the same workflow challenges that the platform was intended to address.

As a next development step for the platform, operative case-based evaluations from multiple evaluators will be combined into self-assembling consensus evaluations with the potential for sophisticated analysis of agreement/disagreement situations. The platform can present a coherent summary of all the recent evaluations as a means to improve the quality of more summative evaluation types (i.e., end-of-rotation, ACGME Milestones) even when this requires that

the performance data be aggregated and structured in specific ways. Currently, we are also helping faculty incorporate custom procedure-specific evaluations, targeted at important procedural steps in common case types. With this in mind, the ability to deliver multiple assessment types can be leveraged in future studies directed at the validation of newly-developed assessment instrument types.

REFERENCES

1. Anderson PA. Giving feedback on clinical skills: are we starving our young? *J Grad Med Educ*. 2012;4:154–158.
2. Williams RG, Verhulst S, Colliver JA, Sanfey H, Chen X, Dunnington GL. A template for reliable assessment of resident operative performance: assessment intervals, numbers of cases and raters. *Surgery*. 2012;152: 517–524. https://doi.org/10.1016/j.surg.2012.07.004. discussion 524-7 Epub 2012 Aug 28.
3. Dougherty P, Kasten SJ, Reynolds RK, Prince ME, Lypson ML. Intraoperative assessment of residents. *J Grad Med Educ*. 2013;5:333–334. https://doi.org/10.4300/JGME-D-13-00074.1.
4. 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;269:377–382. https://doi.org/10.1097/SLA.0000000000002554.
5. Fryer JP, Teitelbaum EN, George BC, et al. Effect of ongoing assessment of resident operative autonomy on the operating room environment. *J Surg Educ*. 2018;75:333–343. https://doi.org/10.1016/j.jsurg.2016.11.018. Epub 2017 Mar 28.
6. Bohnen 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:e118–e130.
7. ACGME Common Program Requirements, https://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf
8. Williams RG, Kim MJ, Dunnington GL. Practice guidelines for operative performance assessments. *Ann Surg*. 2016;264:934–948.
9. 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:e90–e96.
10. Larson JL, Williams RG, Ketchum J, et al. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 2005;138:640–649.
11. Thanawala R, Jesneck J, Seymour NE. Novel educational information management platform improves the surgical skill evaluation process of surgical residents. *J Surg Educ*. 2018;75:e204–e211.
12. Firefly Medical Education Platform, https://www.fireflylab.org
13. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87:1401–1407.
14. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 2005;138:640–647. discussion 647-9.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2019.08.017.