Assessing Interrater Reliability of a Faculty-Provided Feedback Rating Instrument

Daniel P. Walsh*, Michael J. Chen*, Lauren K. Buhl, Sara E. Neves and John D. Mitchell

Department of Anesthesiology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA.
*Co-first authors (Daniel P. Walsh and Michael J. Chen).

ABSTRACT: High quality feedback on resident clinical performance is pivotal to growth and development. Therefore, a reliable means of assessing faculty feedback is necessary. A feedback assessment instrument would also allow for appropriate focus of interventions to improve faculty feedback. We piloted an assessment of the interrater reliability of a seven-item feedback rating instrument on faculty educators trained via a three-workshop frame-of-reference training regimen. The rating instrument’s items assessed for the presence or absence of six feedback traits: actionable, behavior focused, detailed, negative feedback, professionalism/communication, and specific; as well as for overall utility of feedback with regard to devising a resident performance improvement plan on an ordinal scale from 1 to 5. Participants completed three cycles consisting of one-hour-long workshops where an instructor led a review of the feedback rating instrument on deidentified feedback comments, followed by participants independently rating a set of 20 deidentified feedback comments, and the study team reviewing the interrater reliability for each feedback rating category to guide future workshops. Comments came from four different anesthesia residency programs in the United States; each set of feedback comments was balanced with respect to utility scores to promote participants’ ability to discriminate between high and low utility comments. On the third and final independent rating exercise, participants achieved moderate or greater interrater reliability on all seven rating categories of a feedback rating instrument using Gwet’s agreement coefficient 1 for the six feedback traits and using intraclass correlation for utility score. This illustrates that when this instrument is utilized by trained, expert educators, reliable assessments of faculty-provided feedback can be made. This rating instrument, with further validity evidence, has the potential to help programs reliably assess both the quality and utility of their feedback, as well as the impact of any educational interventions designed to improve feedback.

KEYWORDS: feedback, interrater reliability, education

Introduction

High quality feedback on resident clinical performance is pivotal to growth and development. Therefore, a reliable means of assessing faculty feedback is necessary. A feedback assessment instrument would also allow for appropriate focus of interventions to improve faculty feedback. Our research group previously systematically developed, tested, refined, and implemented a seven-item feedback rating instrument, as depicted in Table 1. To this point, we have only utilized it as a group where all raters must come to a consensus on ratings and in an automated fashion using machine learning based on group ratings.

Both of those methods have their own merits and concerns. Although rating in a group fashion can lead to insightful discussions in which raters can explain their rationale to one another, it results in a more time-consuming process as raters must coordinate schedules to perform group ratings, and the person-hours required to rate items increases based on the number of raters present. Conversely, while our initial efforts using machine learning to automate ratings using text processing achieved 74.4–82.2% accuracy with an impressive processing capability to complete thousands of ratings in minutes, we still plan to periodically assess small batches of automated ratings against manually-reviewed ratings to safeguard against drift and potentially retrain models.

One way to address these concerns would be to have individual raters use the seven-item rating instrument; however, this approach depends on reasonable inter-rater reliability (IRR). In order to evaluate our rating instrument for use by independent raters, we piloted an assessment of its IRR following frame-of-reference training with experienced faculty educators. Our hypothesis was that by the end of a training regimen of three, one hour-long workshops and three independent rating exercises, participants would achieve moderate or greater inter-rater reliability on all rating categories.

Methods

Our research group previously reported a seven-item feedback rating instrument developed by three anesthesiologists, each with over five years of educational experience. Users must first assess the presence or absence of six predefined feedback traits: actionable, behavior focused, detailed, negative feedback, professionalism/communication, and specific; then, they must assess the overall utility of feedback with regard to devising a resident performance improvement plan on an ordinal scale from 1 to 5. Definitions and examples for each rating item are displayed in Table 1. This rating instrument was previously used by the creators to rate 1925 feedback comments collected over the course of roughly 8 months (October 2013 to July 2014) from anesthesiology residency programs at 4 institutions.
Table 1. Definitions for the six binary feedback traits and utility with emblematic examples.

| TERM                      | DEFINITION                                           | EXAMPLES                                                                 |
|---------------------------|------------------------------------------------------|--------------------------------------------------------------------------|
| Actionable                | Identifies areas for residents to work on improving. | “Take advantage of this rotation to become more familiar with fiberoptic use.” |
| Behavior Focused          | Notes something done by the resident as modifiable or changeable. Raters differentiated the definition of a behavior from a characteristic, or personality attribute of the resident. | Behavior: “The resident placed arterial lines with enthusiasm.” Characteristic: “The resident was enthusiastic.” |
| Detailed                  | Provides ample information describing observed cases or actions which occurred, but not necessarily how a resident performed. | “We were assigned to three transcatheter aortic valve replacement cases today, one of which was a valve-in-valve procedure.” |
| Negative Feedback         | Notes areas the resident could improve on. Does not necessarily have to be hurtful or personal. | “The resident had difficulty identifying areas of bronchial anatomy during fiberoptic intubation.” |
| Professionalism / Communication | Notes an exceptional level of planning, preparation, and/or communication—or a lack of such. | “The resident actively engaged with nursing and surgery to confirm fluid management strategies, and ensured all supplies were organized to facilitate rapid sequence induction.” “The resident did not ensure that other sizes of double lumen tubes were available before bringing the patient back.” |
| Specific                  | Provides information related to the resident’s actions. | “The resident was able to easily switch to a ‘through and through’ approach after their first arterial line placement attempt was unsuccessful.” |
| Utility                   | Assessment of whether feedback can help devise a performance improvement plan for the resident. | High-Utility Example: Did a great job, just needs to work on thinking and doing things quickly on their feet. For example, after we put in the LMA in our converted MAC, the patient was coughing and not tolerating the LMA. Rather than use their drawn up syringe of propofol and giving a bolus dose, they used the pump to try and deliver a bolus dose which took slightly longer. I think with time in their training, they will naturally continue to improve in this area.” |

Edited from table in Using Machine Learning to Evaluate Attending Feedback on Resident Performance by Neves SE, Chen M, et al. in Anesth Analg.

Examples provided for Actionable, Behavior Focused, Detailed, Negative Feedback, Professionalism / Communication, and Specific are synthetic examples created to exemplify statements containing their respective feedback traits. The example for utility is a genuine, deidentified feedback comment left by faculty on a resident’s performance which achieved the maximum utility score (5 out of 5); this example was also edited to use gender neutral pronouns.

The Actionable, Behavior Focused, Detailed, Negative Feedback, Professionalism / Communication, and Specific rating categories were treated as binary items, with raters noting whether they believed a given comment was emblematic of the respective categories (true) or not (false).

Beth Israel Deaconess Medical Center (BIDMC), University of Rochester Medical Center, University of Kentucky College of Medicine, and University of California San Diego. Feedback comments were qualitative free text entered into an electronic evaluation system by an attending physician who worked with a resident on a single day or over a short period of time.

The study was conducted under approval by BIDMC’s Institutional Review Board (IRB), the Committee on Clinical Investigation, under protocol #2011P000018 (Improving Feedback for Residents via Faculty Education); study participants were verbally consented. Participating raters examined in this inter-rater reliability assessment are co-investigators on the study. In order to coordinate schedules to avoid having a relatively equal distribution of utility scores as depicted in Figure 1, but otherwise be randomly constructed from the dataset. Feedback Set A was used for training purposes to introduce raters to the rating instrument; as there were no independent ratings for Set A it is not included in the interrater reliability analysis. Feedback Sets B, C, and D were used for independent rating exercises; Sets B and C were also reviewed during subsequent teleconference workshops for further review.

Figure 2 depicts an overview of how these sets were used to train participants on the feedback rating instrument via a series of three 60-minute teleconference workshops led by an original creator of the rating instrument and three independent rating
exercises. Participation began on October 7, 2020 with the initial workshop and ended on December 14, 2020 after participants provided ratings on the final independent feedback rating set. Participants were provided a reference guide on how to use the instrument with definitions and examples of each rating category (Table 1); informed that the Actionable, Behavior Focused, Detailed, Negative Feedback, Professionalism / Communication, and Specific rating categories were to be treated as binary items with raters noting whether they believed a given comment was emblematic of the respective categories (true) or not (false); and that for Utility they were to assign integer scores of 1 to 5 based on an ordinal scale where scores of 1 signified low utility and scores of 5 signified high utility. The original raters’ group consensus ratings would serve as the gold standard for participants to compare their ratings to after their independent rating attempts.

In the first workshop, participants were introduced to the rating instrument and asked how they would rate the feedback examples using the rating instrument; discussion was encouraged when ratings differed between members and/or the instructor to assess rationale. Afterwards, participants were assigned a set of feedback examples to independently rate and asked to return their ratings to the study team within one week. The IRR for the independent rating exercise was measured and reported to the instructor to identify areas with the lowest IRR to focus on in subsequent training sessions. In the subsequent workshop, the instructor led participants in a review of their ratings on their recent rating exercise while focusing on the identified areas. Participants were then given the next feedback set to rate and return within one week. This cycle of group workshops and independent rating exercises was repeated until participants had completed a total of three workshops and three rating exercises, as depicted in Figure 2. After each workshop, the original creators’ ratings were shared with participants for any feedback examples which were not reviewed during the call due to time constraints. As there was no review workshop following the third and final independent rating exercise, participants received the original creators’ ratings for the full set of feedback examples. All participants provided ratings for all examples; there were no missing data and no loss to follow-up.

The IRR for each feedback trait was measured using R5,6 to calculate Gwet’s first-order agreement coefficient (Gwet’s AC1)7 and interpreted using Landis & Koch’s rule of thumb.8 Therefore, Gwet’s AC1 values of < 0.00, 0.00 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and 0.81 to 100 indicate poor, slight, fair, moderate, substantial, and almost perfect agreement levels respectively. Unlike Cohen’s Kappa,9 Gwet’s AC1 allows for chance-correlated agreement calculations between more than two raters and is robust against a “Kappa paradox” where low kappa values are reported.

Figure 1. Distribution of feedback examples with respect to original raters’ utility scores.

*Feedback Set A consisted of 20 feedback comments, and was used for training purposes to introduce raters to the rating instrument.

*Feedback Sets B, C, and D each consisted of 20 feedback comments, and were used for independent rating exercises. Sets B and C were also reviewed during subsequent teleconference workshops for further review.
despite high observed agreement values. The IRR for utility scores was measured using R$^5$, to calculate intraclass correlation (ICC), specifically two-way random effects, consistency, multiple raters/measurements. ICC values were interpreted using Koo & Li’s guidelines so that values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 indicate poor, moderate, good, and excellent reliability, respectively. As the interpretation thresholds for Kappa values such as Gwet’s AC1 are different from those of ICC, we decided to use their respective thresholds of 0.41 or greater and 0.5 or greater to determine whether “moderate” IRR was achieved on all rating categories.

**Results**
The distribution of feedback examples with respect to original raters’ utility scores are displayed in Figure 1. While the full feedback set ($n = 1925$) was originally highly skewed towards low utility with a skewness of 1.10 and mean utility rating of $1.94 \pm 1.19$, utility for sets A – D ($n = 80$) had a mean of $3.00 \pm 1.42$ and skewness of 0, while sets B – D ($n = 60$) had a
mean utility of 2.95 ± 1.49 and skewness of 0.09. All three participants attended all three workshops for the entire duration and returned each rating exercise within seven days. The training took a total of nine weeks starting from when participants attended the first workshop to when the final independent rating exercises were returned.

On the final, third rating exercise, all rating categories achieved at least moderate IRR: participants achieved near-perfect IRR on two feedback traits with Gwet’s AC1 values of 0.88 for actionable and 0.83 for professionalism / communication; substantial IRR on three feedback traits with Gwet’s AC1 values of 0.80 for behavior focused, 0.74 for detailed, and 0.67 for specific; and moderate IRR for negative feedback with a Gwet’s AC1 value of 0.54. For utility score, an ICC value of 0.90 (95% CI of 0.80 - 0.96) was achieved across the final exercise. As Koo & Li’s guidelines do not note how to handle scenarios where ICC is on a threshold value, we classified set C’s ICC as “good-excellent” to represent the tiers the value straddles.13 Tables 2 and 3 report IRR and agreement rates for each rating category on each rating exercise. Overall, IRR on the final rating exercise increased compared to the initial rating exercise, with higher Gwet’s AC1 scores on four of seven categories: actionable, detailed, professionalism / communication, and specific. The latter two categories showed the greatest improvement by the final rating exercise, as professionalism / communication IRR improved by three tiers from fair to almost perfect (Gwet’s AC1 of 0.29 to 0.83), and specific IRR improved by three tiers from slight to substantial (Gwet’s AC1 of 0.20 to 0.67). Conversely, IRR for three categories actually dropped by one tier from the first to the final rating exercise: Behavior Focused (Gwet’s AC1 of 0.83 to 0.80), Negative Feedback (Gwet’s AC1 of 0.68 to 0.54), and Utility (ICC of 0.95 to 0.90).

As depicted in Table 3, participants’ utility score ratings generally exhibited adjacent agreement (ie, all participants’ ratings for an example stayed within one point of each other) with 82% adjacent agreement across all 60 feedback examples. Of the 11 feedback examples on which adjacent agreement was not observed, only two examples showed differences of three points between participants’ ratings. Both of these examples came from Set C, which had the worst ICC of all the sets, contained a total of nine examples where adjacent agreement was not observed, and displayed worse IRR on the actionable, behavior-focused, and negative feedback traits compared to the prior feedback set B.

**Discussion**

Through the frame-of-reference training regimen detailed in Figure 2, we were able to train the three participating anesthesia faculty on the seven-item feedback rating instrument. On the final rating exercise, moderate IRR was achieved for all seven rating items. Furthermore, high IRR was achieved on six of the seven rating items: actionable, behavior focused, detailed, professionalism / communication, specific, and utility. The results of this pilot indicate that our hypothesis was correct, as by the end of a training regimen of three 60-minute workshops and three independent rating exercises, participants did

---

### Table 2. Table of inter-rater reliability and percent agreement for feedback trait ratingsa.

| SET B (N = 20) | SET C (N = 20) | SET D (N = 20) |
|----------------|---------------|---------------|
|                | Gwet’s AC1b   | Percent       | Gwet’s AC1c   | Percent       | Gwet’s AC1b   | Percent       |
|                |               | Agreementd    |               | Agreementd    |               | Agreementd    |
| Actionable     | 0.68 (Substantial) | 83%          | 0.56 (Moderate) | 73%          | 0.88 (Almost Perfect) | 93%          |
| Behavior       | 0.83 (Almost Perfect) | 87%          | 0.80 (Substantial) | 83%          | 0.80 (Substantial) | 83%          |
| Focused        | 0.54 (Moderate)  | 77%          | 0.81 (Almost Perfect) | 90%          | 0.74 (Substantial) | 87%          |
| Detailed       | 0.68 (Substantial) | 83%          | 0.47 (Moderate)  | 73%          | 0.54 (Moderate)  | 77%          |
| Negative       | 0.29 (Fair)     | 63%          | 0.55 (Moderate)  | 77%          | 0.83 (Almost Perfect) | 87%          |
| Feedback       | 0.35 (Fair)     | 60%          | 0.35 (Fair)      | 67%          | 0.67 (Substantial) | 83%          |
| ProfComm       | 0.29 (Fair)     | 60%          | 0.35 (Fair)      | 67%          | 0.67 (Substantial) | 83%          |
| Specific       | 0.35 (Fair)     | 60%          | 0.35 (Fair)      | 67%          | 0.67 (Substantial) | 83%          |

---

*Set A was used to introduce the rating instrument and train participants on its use, and therefore was not graded.

1Gwet’s AC1: Gwet’s first-order agreement coefficient. 2Gwet’s AC1 was calculated using R. Results for Gwet’s AC1 are reported with the value followed by interpretation in parentheses per Landis & Koch’s interpretations for kappa values and strength of agreement, and are likewise color-coded as follows: < 0.00 is poor (unused), 0.00 to 0.20 is slight agreement (red), 0.21 to 0.40 is fair (orange), 0.41 to 0.60 is moderate (yellow), 0.61 to 0.80 is substantial (blue), and 0.81 to 1.00 is almost perfect (green and bold font).

3Percent Agreement represents the percent of comments on which all three participants unanimously agreed on whether the feedback trait was present in a comment or not.

4ProfComm: Abbreviation for Professionalism / Communication trait.
in fact achieve moderate or greater interrater reliability on all rating categories.

The markedly improved IRR of the professionalism / communication and specific categories was likely due to discussions held in subsequent review workshops where the participants and instructor were able to explain their rationales. These discussions mirrored the original raters’ experiences in piloting this rating instrument with group consensus ratings, as iterative discussions on how to categorize feedback examples helped the group refine definitions. For study participants, these discussions also likely contributed to the improved ICC for utility between set C (0.75) and Set D (0.90) and improved adjacent agreement (55% to 95%). Participants were able to adjust their understandings of category definitions as their colleagues explained what aspects of each comment influenced their decision making. Although we were able to achieve moderate IRR on all categories with independent raters, we plan to incorporate occasional double rating or periodic group rating. A mixed approach such as this enables us to process the vast amount of feedback generated in a timely fashion, provides a way to confirm raters are still achieving moderate IRR or better with their ratings, and still allows for insightful discussions between raters to share their rationales behind ratings with one another.

Utility is arguably the central category in this rating instrument, with the presence or absence of the other categories contributing to the overall utility score. Artificially balancing feedback sets with respect to utility could have provided an inauthentic sample compared to what participants will likely observe when rating future feedback batches. It is also possible that informing participants that the feedback sets they would be working with were balanced with respect to utility tempered their expectations and led them to create balanced ratings in turn. However, we believe letting participants experience a balanced spectrum of utility ratings for the purpose of training improved their ability to discriminate high utility feedback from low utility feedback. Additionally, during workshops the instructor alerted participants that the vast majority of feedback they would review in the future would tend to be on the lower end of utility scores.

We decided to use only participants from one institution to facilitate training logistics, but this choice precluded us from exploring the rating instrument’s external validity with respect to raters. The inclusion of raters from multiple institutions is a future area for investigation as we move towards rating newer batches of feedback; however, the fact that the feedback examples used to train the raters in the present study came from four different institutions, which collectively represent three of four US Census Bureau regions, speaks to the population validity of our results. Multi-institutional training sessions could potentially be facilitated via pre-recorded modules and other on-demand materials. While our pilot’s participants consisted exclusively of only clinician educators who work in residency education, we believe this to be the ideal target population for use of this faculty-provided feedback rating instrument.

Warm et al. created a similar “Feedback on Feedback” instrument designed to rate faculty on the feedback they provided to participants. The independent ratings of participants were followed by the 95% confidence interval. Interpretations for ICC values taken from Koo & Li’s guidelines, where values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. Calculations performed using R. ICC calculations only took into account the independent ratings of participants.

### Table 3. Table of intraclass correlation, agreement rates, and mean scores for utility score ratings.

|                          | SET B (N = 20) | SET C (N = 20) | SET D (N = 20) |
|--------------------------|----------------|----------------|----------------|
| ICC - Participants Onlyb | 0.95 (Excellent), 95% C.I. [0.88 - 0.98] | 0.75 (Moderate-Good),c 95% C.I. [0.46 - 0.89] | 0.90 (Good-Excellent),d 95% C.I. [0.80 - 0.96] |
| Adjacent Agreementa      | 95%            | 55%            | 95%            |
| Original Mean Utility Score ± SDg | 2.55 ± 1.36 | 3.30 ± 1.30 | 3.00 ± 1.45 |
| Participant 1 Mean Utility Score ± SDg | 2.80 ± 1.15 | 2.95 ± 0.83 | 2.65 ± 0.93 |
| Participant 2 Mean Utility Score ± SDg | 2.40 ± 1.31 | 2.80 ± 1.20 | 3.00 ± 1.12 |
| Participant 3 Mean Utility Score ± SDg | 2.55 ± 1.32 | 3.75 ± 0.97 | 3.10 ± 0.91 |
| All Participants Mean Utility Score ± SDg | 2.58 ± 1.25 | 3.17 ± 1.08 | 2.92 ± 1.00 |

*a* Set A was used to introduce the rating instrument and train participants on its use, and therefore was not graded. Its mean utility score from the original raters was 3.15 with a standard deviation of 1.57.

*b* ICC: Intraclass correlation - Two-way random effects, consistency, multiple raters/measurements.13 Reported as the calculated ICC value with interpretation in parentheses followed by the 95% confidence interval. Interpretations for ICC values taken from Koo & Li’s guidelines, where values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. Calculations performed using R. ICC calculations only took into account the independent ratings of participants.

*c* Set C’s ICC was measured to be 0.745, which rounds to 0.75. As Koo & Li’s guidelines do not note how to handle scenarios where ICC is on a threshold value, we classified set C’s ICC as “moderate-good” to represent the tiers the value straddles.14

*d* Set D’s ICC was measured to be 0.904, which rounds to 0.90. As Koo & Li’s guidelines do not note how to handle scenarios where ICC is on a threshold value, we classified set D’s ICC as “good-excellent” to represent the tiers the value straddles.14

*e* Adjacent Agreement represents the percent of comments on which all three participants in fact achieve moderate or greater interrater reliability on all rating categories.

In the table, ICC values are rounded to two decimal places for clarity. The values reported for mean utility score note the mean value of all ratings within a given feedback set, with the corresponding standard deviation reported next to it.
their residents. Although their four assessment criteria differed from ours and used a 1 to 5 ordinal scale for each assessment, both instruments place value on feedback providing details on observed practices and described specific areas of strengths/weaknesses. Both instruments also assess whether the feedback being provided would ultimately promote improvement in clinical behavior. The similarities between our independently developed instruments and common end-goals of using feedback assessments for targeted faculty development initiatives provide insight into how highly educators value useful feedback. Unfortunately, Wann et al. did not provide details on whether their feedback reviewers participated in any training regimens to establish consensus for distinguishing rating levels from one another, nor do they comment on any potential interrater reliability analyses between their reviewers.

Although our pilot focused on assessing inter-rater reliability of assessors reviewing feedback left from faculty, we recognize that it does not directly address the potential variability in assessments from the faculty leaving the feedback. Gingerich et al. discuss that variability in assessments can stem from multiple non-mutually exclusive origins: assessors may not be using assessment criteria correctly; assessments may vary due to human judgment being innately imperfect and subjective to biases; and/or assessors may interpret observations differently due to potentially meaningful idiosyncrasies in assessors’ cognition. Therefore, it is crucial to first establish a method to systematically assess faculty-provided feedback in order to detect for outliers and to assess potential interventions which aim to address assessment variability that stem from incorrect application of assessment criteria, biases, and other undesirable causes.

While this training regimen did result in participants achieving at least moderate IRR on all rating categories, we identified areas for improvement. For example, workshops were scheduled for one hour for convenience, and any feedback examples not covered during the workshops were sent to participants with their original ratings for self-review. Therefore, it may be better to lengthen future workshops to allow for discussion of each rating category for all examples. Additionally, the decision to spend time during workshops focusing on the rating categories with low IRR came at the expense of not fully addressing the other categories. This approach likely contributed to set B’s three highest IRR categories each dropping by one IRR tier on the subsequent set. Spending time reinforcing good performance is a fundamental part of education, and lengthening future workshops will afford more flexibility to do this by letting participants review more examples in all rating categories to prevent IRR from drifting downward.

In addition to potentially increasing workshop duration, we are also considering providing more materials for participants to refer to when rating feedback. For example, providing a set of slides with in-depth rating rationales for feedback examples could help guide raters in identifying aspects of feedback that influence each rating category. Having more worked examples can help raters understand what the “boundary” should be for deciding whether a feedback trait is present or absent. Such an approach would be particularly useful for the negative feedback trait as participants noted they were often uncertain on how to score it. Additionally, providing reference materials would be useful for refreshing previously trained raters who may not have rated feedback recently. Alternatively, having pre-recorded videos on demand featuring multiple instructors or trained raters dissecting feedback examples could help prospective raters who prefer audiovisual media. This additional media could also serve to supplement training workshops as preparatory work or even replace training workshops entirely, addressing many logistical concerns.

So far, this instrument has only been used in the context of attending anesthesiologists involved in residency program administration reviewing feedback on anesthesia residents; however, the rating instrument’s categories are not limited to that context. This rating instrument could be adapted for use in other specialties and for other providers. These kinds of investigations will become more feasible as we continue to train independent raters and refine our machine learning models to increase throughput potential.

Conclusion
At the end of a series of training sessions and rating exercises, participants achieved high IRR on six of seven rating categories (actionable, behavior focused, detailed, professionalism/communication, specific, and utility) and moderate IRR on the remaining category (negative feedback). This pilot illustrates that when this instrument is utilized by trained clinician educators who work in resident education, reliable assessments of faculty-provided feedback can be made. With further validity evidence, this rating instrument has the potential to help programs reliably assess both the quality and utility of their feedback, as well as the impact of any educational interventions designed to improve feedback.

Acknowledgements
Stephanie Jones, MD & Cindy Ku, MD – Co-creators of rating instrument.

Faculty and residents from: Beth Israel Deaconess Medical Center, University of Rochester Medical Center, University of Kentucky College of Medicine, & University of California San Diego.

Declaration of Interest Statement
No potential competing interest was reported by the authors.

The Authors declare that there is no conflict of interest.

Ethical Approval
Not applicable, because this article does not contain any studies with human or animal subjects.
Informed Consent
Not applicable, because this article does not contain any studies with human or animal subjects.

Trial Registration
Not applicable, because this article does not contain any clinical trials.

ORCID iD
Michael J. Chen. https://orcid.org/0000-0001-8185-5163

REFERENCES
1. Ende J. Feedback in clinical medical education. JAMA. 1983;250(6):777–781. doi:10.1001/jama.1983.03340060055026. PMID: 6876333.
2. Mitchell JD, Ku C, Diachun CAB, et al. Enhancing feedback on professionalism and communication skills in anesthesia residency programs. Anesth Analg. 2017;125(2):620–631. doi:10.1213/ANE.0000000000002143
3. Neves SE, Chen MJ, Ku CM, et al. Using machine learning to evaluate attending feedback on resident performance. Anesth Analg. 2021;132(2):545–555. doi:10.1213/ANE.000000000005265
4. von Elm E, Altman DG, Egger M, et al. The strengthening of reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies [published correction appears in Ann Intern Med. Ann Intern Med. 2007;147(8):573-577. doi:10.7326/0003-4819-147-8-200710160-00010
5. R Core Team. R: a language and environment for statistical computing [computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. R Core Team. R: a language and environment for statistical computing [computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. Gwet KL. irrCAC: computing chance-corrected agreement coefficients (CAC) [computer software]. R package version 1.0. 2017. https://CRAN.R-project.org/package=irrCAC
7. Gwet KL. Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. Gaithersburg, 4th edn. STATAxis Publishing Company; 2014.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. doi:10.2307/2529310
9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure. 1960;20(1):37-46. doi: 10.1177/0013164460020000104
10. Zee S, Soriano N, Comeorto R, Baldi I. High agreement and high prevalence: the paradox of Cohen’s kappa. Open Nurs J. 2017;11(3):211–218. Published 2017 Oct 31. doi:10.2174/1874434601711010211
11. Garner M, Lemon J, Singh IFFP. IRR: various coefficients of interrater reliability and agreement. R package version 0.84.1. 2019. https://CRAN.R-project.org/package=IRR
12. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol. 1996;1(1):30-46. doi:10.1037/1082-989X.1.1.30
13. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research [published correction appears. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
14. Warm E, Kelleher M, Kinnear B, Sall D. Feedback on feedback as a faculty development tool. J Grad Med Educ. 2018;10(3):354-355. doi:10.4300/JGME-D-17-00876.1
15. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the ‘black box’ differently: assessor cognition from three research perspectives. Med Educ. 2014;48(11):1055-1068. doi:10.1111/medu.12546