Faculty Perceptions of Frame of Reference Training to Improve Workplace-Based Assessment

Jennifer R. Kogan, MD
Lisa N. Conforti, MPH
Eric S. Holmboe, MD

ABSTRACT

Background Workplace-based assessment (WBA) is a key assessment strategy in competency-based medical education. However, its full potential has not been actualized secondary to concerns with reliability, validity, and accuracy. Frame of reference training (FORT), a rater training technique that helps assessors distinguish between learner performance levels, can improve the accuracy and reliability of WBA, but the effect size is variable. Understanding FORT benefits and challenges help improve this rater training technique.

Objective To explore faculty's perceptions of the benefits and challenges associated with FORT.

Methods Subjects were internal medicine and family medicine physicians (n=41) who participated in a rater training intervention in 2018 consisting of in-person FORT followed by asynchronous online spaced learning. We assessed participants' perceptions of FORT in post-workshop focus groups and an end-of-study survey. Focus groups and survey free text responses were coded using thematic analysis.

Results All subjects participated in 1 of 4 focus groups and completed the survey. Four benefits of FORT were identified: (1) opportunity to apply skills frameworks via deliberate practice; (2) demonstration of the importance of certain evidence-based clinical skills; (3) practice that improved the ability to discriminate between resident skill levels; and (4) highlighting the importance of direct observation and the dangers using proxy information in assessment. Challenges included time constraints and task repetitiveness.

Conclusions Participants believe that FORT training serves multiple purposes, including helping them distinguish between learner skill levels while demonstrating the impact of evidence-based clinical skills and the importance of direct observation.

Introduction

Workplace-based assessment (WBA) is a key assessment strategy in medical education, particularly in competency-based medical education. Improving WBA quality requires faculty development,[1-4] which is necessary to improve faculty's ability to observe, synthesize observations into a judgment, encode judgments into an entrustment rating, provide feedback, and coach learners.[3-5] However, the impact of faculty development interventions on the reliability and accuracy of WBA in medical education has been small to negligible.[6-12] As such, more work is needed to delineate the unique effects of the various components and implementation designs of rater training.[4,5]

Performance dimension training (PDT) and frame of reference training (FORT) are 2 rater training techniques that can improve performance appraisal assessments.[13,14] PDT trains assessors to recognize the appropriate behaviors or dimensions of a given competency or skill using evidence-informed definitions supported by examples using written vignettes, videos, or role-plays.[13,14] FORT helps assessors discriminate between variations in the quality of demonstrated skills by having participants individually assess multiple videos of individual learners with different skill levels and then together discuss discrepancies in observations and ratings.[13,14] For example, during PDT, assessors are asked to identify the behaviors that constitute aspirational shared decision-making and are shown examples of these behaviors in video vignettes. Then, during FORT, facilitators ask team members to individually assess several stimulus videos of a resident counseling a patient about starting a statin, in which the resident demonstrates a range of skills from poor to aspirational. Assessors then discuss the videos as a group using a compare-contrast approach, sharing assessments and discussing discrepancies to create a shared mental model to guide assessment judgements.[13,14] In a business setting, PDT and FORT can improve assessment accuracy and minimize rating variability by decreasing rater errors or biases unrelated to the targeted performance behaviors.[13,14] The impact of

DOI: http://dx.doi.org/10.4300/JGME-D-22-00287.1

Editor's Note: The online version of this article contains an overview of creation of the training videos and expert answer keys, components of rater training faculty development, the focus group interview guide, and the survey used in the study.

these techniques in medical education have been more variable.[9-12]

We previously published a rater training study that incorporated emerging rater cognition theories (the use of variable frames of reference, inference, and uncertainty in how to translate observations into numerical ratings) to determine which components of rater training might improve WBA.[15-19] In that study we explored how PDT affected participants' approach to WBA.[19] More recently, we demonstrated that rater training which included FORT increased the accuracy and specificity of observations.[20] Clarifying the mechanisms of FORT in medical education may provide further insights into rater cognition and how best to conduct FORT to improve WBA quality and accuracy.[21,22] To our knowledge there has not been a study examining faculty perceptions of FORT in medical education and the mechanism by which it influences assessors' judgments.

The purpose of this study was to understand assessors' perceptions of FORT, in particular its benefits and challenges.

Objectives
To explore faculty's perceptions of the benefits and challenges associated with frame of reference rater training.

Findings
Participants felt frame of reference training offered an opportunity to apply skills frameworks via deliberate practice, demonstrated the importance of evidence-based clinical skills, improved their ability to discriminate between resident skill levels, and highlighted the importance of direct observation.

Limitations
It is uncertain how findings generalize to other specialties or rater training for other skills.

Bottom Line
This study provides an understanding for how frame of reference training impacts faculty's beliefs about and approach to workplace-based assessment.

Methods

We previously published a randomized controlled trial of a longitudinal rater training intervention to improve WBA.[20] This current study focuses only on the intervention group of that study and on the perceptions of FORT.

Participants

Between December 2017 and August 2018, we emailed family and internal medicine residency program directors at 138 programs in 6 Midwest states and 186 programs in 5 Mid-Atlantic states soliciting their interest to enroll faculty in the study. Program directors provided email addresses for potential participants. Eligible participants needed to be general practitioner faculty who (1) were responsible for outpatient clinical training and evaluation of residents; (2) provided outpatient care for their own patient panel; (3) held a faculty position for at least 1 year; and (4) were available for a 2-day study session. Participants who agreed to enroll were initially randomized to the control (n=45) or intervention arm (n=49). This current study focuses only on the 41 intervention group participants who completed the study (7 dropped out post-randomization and 1 dropped out mid-trial secondary to illness). The control group participants are not included in the current study. Participants received a modest $150 honorarium from the Accreditation Council for Graduate Medical Education. The intervention group was eligible for up to 14.25 Continuing Medical Education and Maintenance of Certification credits.

Development of Stimulus Videos

Between June 2016 and June 2018, with the assistance of 6 experts in physician-patient communication and trainee assessment, and using evidence from the literature, we created stimulus videos depicting residents taking a history from or counseling patients.[23-28] We created 27 videos (9 scenarios, each demonstrating 3 different levels of resident skill) for rater training using recommended guidelines.[29] In the online supplementary data, we summarize the steps we used to create the videos and their associated answer keys with the expert-informed consensus entrustment rating and narrative assessment.

Intervention and Assessments

At baseline, participants completed a self-administered demographic web-based questionnaire and assessed 10 stimulus videos (5 history taking and 5 counseling) using an online rater assessment form asking them to identify what the resident in the video did well, what required improvement, and how they would supervise the resident going forward (prospective entrustment decision).[20] Participants attended 1 of 4 rater trainings in the fall of 2018. The rater trainings were 2-day, in-person, 3-hour workshops that immediately followed the baseline assessment. The rater training content and format (online supplementary data) were informed by our prior research and included PDT and FORT.[9,15-19] During PDT, participants created frameworks of the skills required for history taking and counseling. They then reviewed an evidence-based framework for each skill and revised their framework as needed.[23-28] During FORT participants applied the framework to the 2

remaining stimulus videos in the series, comparing their assessments to the answer keys as a group.

After each of the 4 rater training workshops, 1 of 4 individuals with expertise in focus group facilitation (but otherwise unrelated to the study) led a focus group about the rater training (focus group guide provided as online supplementary data). Focus groups occurred immediately after the rater training so that participants could provide specific feedback about the rater training. Study investigators were not present during the focus groups, which were audio-recorded, transcribed, and de-identified.

Four weeks after the in-person workshops, participants started 3 asynchronous, online, spaced learning FORT modules with timed deliverables spaced 6 weeks apart using the Canvas Learning Management System (Instructure Inc). In spaced learning, a course is divided into short duration modules with breaks between the sessions. We included spaced learning because knowledge retention is enhanced when learning sessions are spaced in time.[30] During each spaced learning module, participants were prompted to watch a series of history taking and counseling stimulus videos, each depicting 2 different levels of resident skill. We instructed participants to use their frameworks to guide observations. After rating each video, participants reviewed the answer key and identified similarities and differences between their own observations and those of the expert. A third video for each series was available for optional review (online supplementary data). A study investigator (J.K. or E.H.) moderated the spaced learning discussion boards.

At a minimum of 4 weeks after the last spaced learning module (March to May 2019), participants watched and rated 10 stimulus videos (5 history taking and 5 counseling). Participants completed a 19-item end-of-study survey focused on the intervention (online supplementary data). Questions asked about the benefits of spaced learning (rated on a 5-point Likert scale where 1=strongly disagree, 5=strongly agree) and spaced learning timing and work effort (rated on a 3-point scale). There were 3 open-ended questions eliciting (1) strengths of spaced learning to improve skills in direct observation; (2) ways spaced learning could be improved; and (3) barriers to spaced learning participation.

Analysis

We used descriptive statistics to summarize demographic and end-of-study survey data. Two investigators (J.K., L.C.) independently coded focus group transcripts and open-ended survey questions using thematic analysis.[31] The investigators began by familiarizing themselves with the data, coding data relevant to the research question, and generating initial themes. The investigators met multiple times to review, discuss, reconcile, and name the themes identified. Themes were then shared with the third investigator (E.H.) for additional input and clarification.

The Institutional Review Board of the University of Pennsylvania Office of Regulatory Affairs approved this study. All participants provided informed consent.

Results

Table 1 summarizes participant demographics. All participants attended 1 of 4 post-workshop focus groups. From the focus groups we identified 4 themes describing FORT benefits and 6 themes describing challenges. Table 2 provides example quotes for each theme.

Benefits of FORT

First, FORT allowed participants to practice applying their previously created frameworks to different videos. Participants described how watching multiple videos of the same encounter enabled them to apply the frameworks to videos demonstrating a progression of resident skill level. Participants explained how watching the same encounter at 3 different resident skill levels promoted deliberate practice. They explained how writing out their observations required them to commit to their assessments, further facilitating deliberate practice.

Second, watching videos of the same encounter at 3 different resident skill levels helped participants gain clarity about the importance of specific clinical skills. Initially some participants doubted that a particular framework behavior was important and necessary for safe, effective, patient-centered care (for example, asking patients to prioritize their agenda at the beginning of a visit). However, when participants watched the video in which the resident started the encounter with agenda-setting, they were able see why that behavior was beneficial and important. Therefore, seeing aspirational performance helped highlight the value and importance of certain clinical behaviors that may have initially been dismissed as unimportant or minimally important.

Third, seeing the same clinical encounter performed at 3 resident skill levels helped participants discriminate between performance levels. While it was sometimes difficult for participants to distinguish between poor and satisfactory performance (calling for direct and indirect supervision respectively) or between satisfactory and aspirational performance

TABLE 1
Baseline Characteristics of Participants

Characteristic	n (%)
Gender	
Female	25 (61)
Male	16 (39)
Age in years, mean (SD)	43.3 (9.8)
Years post-residency, mean (SD)	11.3 (9.8)
Completed fellowship	9 (22)
Primary specialty	
Internal medicine	28 (68)
Family medicine	13 (32)
Academic rank, n (%)	
Instructor	4 (10)
Assistant professor	18 (44)
Associate professor	10 (24)
Professor	1 (2)
Other/not applicable	8 (20)
Institution type	
University-based	17 (41)
Community-based, university-affiliated	15 (37)
Community-based, non-university-affiliated	8 (20)
Other[a]	1 (2)
Educational leadership roles[b]	
Program, associate, or assistant director	20 (49)
Core faculty	21 (51)
Clinical competency committee chair or member	16 (39)
Other (resident clinic site director, assistant/associate fellowship director, medical school clinical rotation course director, etc)	14 (34)
None	5 (12)
Outpatient half days per week seeing own patients, mean (SD)	3.0 (1.9)
Years precepting residents in outpatient setting, mean (SD)	9.0 (8.1)
Half days per week precepting residents in outpatient setting, mean (SD)	2.9 (1.8)
Precepting supervision practices (typically see all patients when precepting)	
Interns in clinic the first 6 months of internship	36 (88)
Interns in clinic the second 6 months of internship	12 (29)
Second- and third-year residents in clinic	8 (20)
Participated in a faculty development workshop in last 5 years focused on:	
Observing and assessing residents in a clinical setting	16 (39)
Competency-based assessment of medical trainees	23 (56)
Giving feedback to medical trainees	29 (71)
Communication skills with patients	17 (41)

[a] Hospital, but not residency, has university affiliation.
[b] Participants were able to select more than one role.

(no supervision needed), watching the sequence of 3 videos helped participants better distinguish between resident skill levels. Participants described how comparing and contrasting residents across the video series enabled them to better understand the range of behaviors for, or variable execution of, a given skill. For example, across the video series, participants could see a range of how much and how well a resident explored the physical, psychological, and emotional impact of a symptom on a patient.

Fourth, participants described how watching the 3-video series emphasized the importance of direct

TABLE 2
Benefits and Challenges Associated with Frame of Reference Training (FORT)

Theme	Example Quote
Benefits of FORT	
Practice applying framework to videos and engaging in deliberate practice	*I think, for me, it felt like this was a chance. . .you've given me this new tool called a framework. I'm not that comfortable with doing it [WBA direct observation]. Let me kind of practice this. And so, I was very much focused more on practicing that skill of how I move between the page [framework] and observing and figuring out how to then record that data*. (Participant [P] 71, Focus Group [FG] 4) FG 2 Moderator: What do you think was the most important piece of all of this? *It was applying the framework to watching the videos across the progression*. (P 11, FG 2) *I liked all the different videos that had the different levels of. . .whether they needed eventually direct or indirect [supervision] and I really liked that. You could relate it to the frameworks and knew everybody's perspectives*. (P 18, FG 2) *The actual repetition of the task in the videos was helpful for me*. (P 82, FG 3) FG 2 Moderator: Tedious. Okay that's fair. What else? What do you guys think about that? *I felt the exact same way. But then I realized that making me focus this much on these little details is what's going to make me see a difference between the subtle differences. The differences between 1 and 3 was huge but 1 and 2 was medium. And 2 and 3 was medium, there's more subtle differences. . .And then I was like okay now I'm understanding the more you write it the more you realize you look for it. So maybe that was perhaps the point of it? The redundancy was the point*. (P 42, FG 2)
Demonstrates importance of specific clinical skills	*I liked the order in which it was done. You started with the intermediate one and without a framework in front of you and just had to judge it like we did in the pretest. And then we got the framework, and you were able to say, "Oh, I saw this. I guess that was what was on the framework." And then when you went back and saw the low and the high one, you're able to very clearly be like. . .especially stuff on the framework you're like, "I don't see how this would be useful. This seems kind of silly." And then you see it in the video and you're like, "Oh, actually I see how that's important." Doing it without the framework, getting the framework and then doing it again was useful for me*. (P 27, FG 3) *I remember when we were doing that exercise too there was one part where we're all going through the history part and like no, that one [referring to an item on the framework] is definitely out. The general consensus among the participants. . .was in relation to [eliciting] the patient priority list [agenda] and that's not necessary. That seems like it's a waste of time. But in the long term, that's the whole goal of all of this. Is to not just get through the visit but long-term success with the patient*. (P 43, FG 3) *It was nice that they defended it [eliciting patient agenda] with actual studies*. (P 42, FG 3)
Helps improve discrimination between performance levels	*I liked how it was the same encounter done differently. . .It was the same person, same encounter. That was the most helpful*. (P 42, FG 2) *When I saw first video, it was medium, I ranked it as a 7. Almost everybody ranked it as a 7. And I could not find—why not 9? Although I couldn't find mistakes. But at the end of the workshop, when I saw worst video and the best video, I knew why it was 7 and not 9*. (P 15, FG 4) *I think the sequentiality of them. We have 3 videos each time and there was development or regression. It gave me a chance to respond. And then you saw the next one, and you were like, "Oh. That was bad, but this is much better." You change your frame of reference. Which I'm sure is part of the point*. (P 69, FG 4) *By the end I was like, "Oh this is the high and this is the mid and this is low." But in the first round I wasn't in the mindset of that. I grew into that and I feel like the reason I grew into it was because they handed us the key. I was like oh I see, I see this one has all the green [indicating more aspirational performance] and this one has most of the red [indicating a lower performance]*. (P 3, FG 2) *I think it was probably valuable to do 3 [videos of same scenario]. I think it still accomplishes its purpose in illustrating, "Look at the behaviors are different." And you're not pointing out, "Oh, this person was less empathetic," you're pointing out, "This person did these things differently and that's what I'm going to point out in terms of. . ." It's good training in terms of how to use this framework*. (P 76, FG 3)

TABLE 2
Benefits and Challenges Associated with Frame of Reference Training (FORT) (continued)

Theme	Example Quote
Emphasizes importance of direct observation and dangers using proxy information	*I found it useful to watch those 3 videos today which were the same clinical scenarios, the same outcome. . .The 3 totally different approaches showed me how much I'm missing by not doing direct observation. Those 3 videos in sequence were very powerful.* (P 5, FG 1) *It was very powerful.* (P 7, FG 1) *Especially when you have the framework. I wrote the same thing, so being able to compare and contrast the poorer and the better interactions*. (P 2, FG 1) *You basically got 3 completely different patient encounters, which could have 3 different outcomes in terms of patient outcomes. You could have gotten the same presentation from all 3 of those encounters. It's kind of like you're getting one presentation, but the reality is that 3 completely different things are happening and that's having serious consequences. The only way to pick up on those potentially serious consequences is to do more direct observation*. (P 62, FG 3) *Watching this same scenario 3 times by 1 person, I realized we could have gotten the same information [during the] presentation at the end. . .But I didn't realize how terrible some [laughter] were at getting the information. Right? They would have come back and presented the same. Boom, boom, boom, boom, boom.* "*And this is what we're going to do.*" *And that would have been correct, I guess. . .But it really makes me realize that I definitely have to do more direct observation with the residents to see how they're obtaining that information.* (P 63, FG 3)
Challenges With FORT	
Similarity between videos in a series	*I will say, the videos run together a little bit*. (P 69, FG 4) *And that's actually a downside. Because as I watched the third video or the second video, I was like, but did they not just do that? Or how much of it was memory from the previous video? The videos were almost too alike*. (P 17, FG 4)
Resident skills too obvious	*But by watching the 3 videos, I had premature closure when I'm like,* "*Oh, this is the bad one. He's going to say a couple of stupid things now.*" *It would be good to have more subtle examples*. (P 76, FG 3) *Maybe the best video should maybe have some qualities, should have more room for improvement. It was a little too good and then the bad guy was a little too bad. I would like them to be a little more mixed so that I could still identify things that. . .because we were really pulling at some of the stuff we said that the good video could do better*. (P 27, FG 3)
Tedium in writing observations	*The ultimate goal was to see the progression. But the redundancy in writing some of the same things you had seen from beginning was to me a little too tedious*. (P 44, FG 2) *It's what you go to do to learn it. Learning is tedious*. (P 11, FG 2)
Groupthink	*I think going along with that and the premature closure, I don't know if it was a good thing or a bad thing, but when we'd see the bad residents, there would be some reactions from everyone. We'd either laugh or whatever. I don't know if that also influenced what we picked up on or what we evaluated as well*. (P 82, FG 3*)* *Yeah, I don't know that I would have noticed the hand holding if it wasn't for everyone who started laughing at it. . .yeah, groupthink*. (P 27, FG3)
Differentiating the resident from the actor	*I wasn't sure whether the person was stiff because they were a poor actor or just not comfortable being a doctor. . .Was I supposed to comment on their poor body language? I don't know. I don't know whether that person had poor body language or was just a bad actor*. (P 2, FG 1) *Yeah, the ones that were looking at the computer more I didn't know if they were reading their script or if they were trying to, you know. . .yeah, I don't know. I agree with that*. (P 6, FG 1)
Removing behaviors from the framework	*We need an objective framework to underlie the art of direct observation. We all need to agree on what that looks like when done well before we can correctly observe it and critique it. I actually would have appreciated some more data on that because they clearly had come up with their framework by a good literature and patient-centered interviewing. We all took some issue with some of the things on that list, and I would have appreciated being confronted with why we're wrong, because I'm sure we are*. (P 76, FG 3) *I didn't love the* "*Circle the things that could be left out of the encounter.*" *I was confused as to what they were getting at and in the end, I was like,* "*So are you getting that off the list?*" *Because essentially we ended with this,* "*Okay, so we don't need that then? But it is evidence-based, so should we do that?*" (P 27, FG 3)

observation. Multiple participants described having an "ah ha" moment when they realized that a resident's oral case presentation might be identical after all 3 video encounters and might not represent what occurred during the patient visit. As such, FORT underscored how a resident's patient presentation was an incomplete and inadequate proxy for what occurred during an office visit. Participants recognized that, while the information a resident obtained from the patient might be the same, the patient's experience during the encounter and subsequent outcome of the visit could substantially differ. This realization further reinforced the value of direct observation.

Challenges With FORT

Participants described several challenges with FORT. Given the similarities between each video in a series, participants described how it was confusing to recall what occurred in each video. Therefore, they questioned whether the same or a different actor should portray the resident in all 3 videos. However, participants also recommended making the differences between the videos more subtle, with the poorly skilled resident a little better and the aspirational resident a little less skilled. Some felt that writing down their observations for each video in the series was tedious, particularly given the similarity of videos in the series. Participants also described how watching videos as a group may have caused cueing when other participants had verbal or non-verbal reactions to a video, potentially promoting groupthink. Participants knew that the residents in the videos were portrayed by actors, so at times they were uncertain if the behaviors they observed should be attributed to the resident's skill or the actor's performance. Finally, during FORT we asked participants if there were any behaviors on the framework that were not essential for safe, effective, patient-centered care. Several participants described their uncertainty taking behaviors off the framework when the frameworks were presented as being evidence-based.

End-of-Survey Results About FORT Spaced Learning

All participants (n=41) completed the end-of-study survey. Participants valued space learning as an approach to improve their skills in direct observation and feedback (TABLE 3) and were favorable regarding the number and timing of spaced learning modules (data not shown).

Almost all participants answered the open-ended questions on benefits of (n=39), areas to improve (n=39), and barriers to participating in (n=38) FORT spaced learning. Survey responses reiterated many of the focus group themes. Participants described how spaced learning afforded them additional opportunity for repeated practice, which helped refresh skills in direct observation while bringing intervening real-world experience to practice. Participants described how repeated practice mitigated losing previously acquired skills, promoting longer-term learning. Practice also reinforced the frameworks, thereby building an "internal model" for the competency being observed. As a result, applying the frameworks started to become "second nature."

Participants also described the benefit of comparing their assessments to the answer keys during spaced learning. The answer keys helped participants differentiate between skill levels across the video series. The combination of seeing more cases and comparing assessments to the answer keys made it easier for participants to differentiate between good and fair resident skill levels. With time, participants described being better able to identify the more subtle differences in resident skills. Additionally, participants valued the ability to compare their assessments to the answer keys to see how they could further improve as evaluators. Finally, the FORT spaced learning continued to serve as a reminder that resident behaviors in the room with a patient may not translate into their oral case presentations.

The greatest challenge with spaced learning was time. Participants described how it was challenging to complete the spaced learning given their competing responsibilities. Some participants wondered if it would be better to assign fewer videos per module or make each video shorter but have more modules. Several commented that typing their observations was tedious. Finally, some participants recommended expanding the modules to skills beyond history taking and physical examination.

Discussion

In this study we explored the mechanism by which FORT, delivered through 2 in-person workshops and 3 asynchronous online spaced learning modules, impacted participants' approach to WBA. We found that FORT provided participants an opportunity to practice applying assessment frameworks, highlighted the importance of specific evidence-based clinical skills in patient care, helped participants improve their ability to discriminate between skill levels, and emphasized the importance of direct observation and the dangers of using proxy information in assessment. There are ongoing calls to improve assessment across the undergraduate to graduate medical education continuum.[32-34] Importantly, the ability to assess

TABLE 3
Faculty Perceptions of Frame of Reference Training Spaced Learning on End-of-Study Survey

Survey Item	Strongly Disagree/Disagree, %	Neutral, %	Agree/Strongly Agree, %	Not Applicable, %
Spaced learning. . .				
was a valuable addition to the in-person training	2.4	17.1	73.2	7.3
helped me increase how often I do direct observation	14.6	22.0	56.1	7.3
improved my skills in direct observation of history taking	7.3	7.3	75.6	9.8
improved my skills in direct observation of counseling	7.3	7.3	75.6	9.8
improved my feedback to learners	7.3	17.1	63.4	12.2
improved my skills in selecting an assessment/entrustment rating	9.8	17.1	63.4	9.8
expert answer keys were a valuable resource	4.9	4.9	75.6	14.6
needed the in-person workshop first	0	0	90.2	9.8

effectively requires training and practice.[2-5,21] To our knowledge this is the first study to explore benefits and challenges of WBA FORT in medical education.

The participants in our study recognized how FORT provided an opportunity to engage in deliberate practice of assessment through repetition, reflection, and feedback using the answer keys. The acquisition of expertise requires deliberate practice,[35] and acquiring expertise in assessment is likely no different. The ability to compare individual assessments to expert raters has previously been identified as an important FORT technique.[13] Deliberate practice requires motivation and endurance, and participants described how this practice sometimes felt tedious and time consuming. As such, more research is needed to determine how best to optimize stimulus video content and delivery.

Faculty evaluation of learners is likely related to faculty's own clinical skills.[15,16] Our findings highlight how FORT may help faculty shift from using their own clinical skills as the standard when evaluating residents. Furthermore, if faculty do not routinely perform a particular skill, or if they do not believe that behavior is important for effective patient care, they are unlikely to comment on it or assess it in their learners. While reviewing evidence-based clinical skills frameworks may suffice to convince a few faculty members of the importance of these specific clinical skills,[19] we found that faculty could better appreciate why certain skills were important and beneficial after seeing the 3-video series, particularly the video of the resident with aspirational skills. Therefore, FORT may highlight the importance of specific clinical skills and the likelihood that faculty would use them in criterion-referenced resident assessment.

Successful implementation of WBA requires that both individuals and the organizational culture value direct observation.[5,36] It can be difficult to get faculty to buy into the importance of direct observation.[37] Watching the 3-video series reinforced the importance of direct observation. This was an unexpected finding. Watching the video series helped participants appreciate, in a more salient and visceral way, how a resident might present a patient the same way despite 3 very different patient encounters. This realization reinforced the importance of direct observation and the dangers and limitations of using proxy information. Going forward, using a video series may be an effective strategy to increase buy-in to direct observation.

The rater training in this study was longer than most previously published studies (6 hours of in-person workshops and 3 online asynchronous spaced learning modules). Effective rater training takes time, and this raises the real challenge of how to integrate more intensive rater training into medical education assessment programs given faculty's limited time and competing responsibilities. That said, if we want to improve assessment, we need to recognize that there are no easy fixes. Effective assessment takes time, practice, and faculty development.[5,21,37] Furthermore, iterative practice is important given that prior rater training interventions have shown drift effects in which assessors show high levels of interrater reliability initially but poor reliability in subsequent performance assessments.[38,39] Successful implementation of WBA requires that individuals and the organizational culture value direct observation.[5,36] While it can be difficult to get faculty to buy into the importance of direct observation, training assessors on resident observation helps establish a sense of trust, reliability, and validity in the feedback that faculty provide to learners after conducting an observation, thereby shifting the cultural view of WBA as time well spent. Future research will need to

explore how best to implement FORT to maximize effectiveness while balancing feasibility.

There are several limitations to our findings. We needed to recruit from multiple residency programs to identify faculty for this study, and the participants' motivation may differ from those who chose not to participate (volunteer bias). Second, it is uncertain how findings generalize to other specialties or rater training for other skills (for example procedural skills). Third, there was variable participation in the spaced learning. Fourth, participants may have described more benefits to training to justify their time participating. Finally, although the impact of WBA is maximized when it is followed by feedback, this study did not address the impact of training on feedback.

Conclusions

Individuals participating in FORT describe how it not only enables deliberate practice to improve discrimination between skill levels but also reinforces the evidence-informed skill frameworks and the importance of direct observation.

References

1. Anderson HL, Kurtz J, West DC. Implementation and use of workplace-based assessment in clinical learning environments: a scoping review. *Acad Med*. 2021;96(suppl 11):164-174. doi:10.1097/ACM.0000000000004366
2. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: the do's, don'ts, and don't knows of direct observation of clinical skills in medical education. *Perspect Med Educ*. 2017;6(5):286-305. doi:10.1007/s40037-017-0376-7
3. Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Adv Health Sci Educ Theory Pract*. 2016;21(2):455-473. doi:10.1007/s10459-015-9614-0
4. Prentice S, Benson J, Kirkpatrick E, Schuwirth L. Workplace-based assessments in postgraduate medical education: a hermeneutic review. *Med Educ*. 2020;54(11):981-992. doi:10.1111/medu.14221
5. Lorwald AC, Lahner FM, Mooser B, et al. Influences on the implementation of Mini-CEX and DOPS for postgraduate medical trainees' learning: a grounded theory study. *Med Teach*. 2019;41(4):448-456. doi:10.1080/0142159X.2018.1497784
6. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examination. *Med Educ*. 1980;14(5):345-349. doi:10.1111/j.1365-2923.1980.tb02379.x
7. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents. *Ann Intern Med*. 1992;117(9):757-765. doi:10.7326/0003-4819-117-9-757
8. George BC, Teitelbaum EN, DaRosa DA, et al. Duration of faculty training needed to ensure reliable OR performance ratings. *J Surg Educ*. 2013;70(6):703-708. doi:10.1016/j.jsurg.2013.06.015
9. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874-881. doi:10.7326/0003-4819-140-11-200406010-00008
10. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized controlled trial. *J Gen Intern Med*. 2009;24(1):74-79. doi:10.1007/s11606-008-0842-3
11. Robertson RL, Vergis A, Gillman LM, Park J. Effect of rater training on the reliability of technical skills assessment: a randomized controlled trial. *Can J Surg*. 2018;61(6):405-411. doi:10.1503/cjs.015917
12. Weitz G, Vinzentius C, Twesten C, Lehnert H, Bonnemeier H, Konig IR. Effects of a rater training on rater accuracy in a physical examination clinical skills assessment. *GMS Z Med Ausbild*. 2014;31(4).doc41. doi:10.3205/zma000933
13. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occupation Org Psychol*. 1994;67(3):189-205. doi:10.1111/j.2044-8325.1994.tb00562.x
14. Feldman M, Lazzara EH, Valderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. 2012;32(4):279-286. doi:10.1002/chp.21156
15. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048-1060. doi:10.1111/j.1365-2923.2011.04025.x
16. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med*. 2010;85(suppl 10):25-28. doi:10.1097/ACM.0b013e3181ed1aa3
17. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Health Sci Educ Theory Pract*. 2013;18(3):325-341. doi:10.1007/s10459-012-9372-1
18. 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

19. Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: a qualitative study. *Med Educ*. 2015;49(7):692-708. doi:10.1111/medu.12733
20. Kogan JR, Dine CJ, Conforti LN, Holmboe ES. Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? A randomized controlled trial [published online ahead of print July 21, 2022]. *Acad Med*. doi:10.1097/ACM.0000000000004819
21. ten Cate O, Balmer DF, Caretta-Weyer H, Hatala R, Hennus MP, West DC. Entrustable professional activities and entrustment decision making: a development and research agenda for the next decade. *Acad Med*. 2021;96(suppl 7):96-104. doi:10.1097/ACM.0000000000004106
22. Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Med Educ*. 2008;42(2):128-133. doi:10.1111/j.1365-2923.2007.02974.x
23. Lane JL, Gottlieb RP. Structured clinical observations: a method to teach clinical skills with limited time and financial resources. *Pediatrics*. 2000;105(4 Part 2):973-977.
24. Makoul G. The SEGUE framework for teaching and assessing communication skills. *Patient Educ Couns*. 2001;45(1):23-34. doi:10.1016/s0738-3991(01)00136-7
25. Lyles JS, Dwamena FC, Lein C, Smith RC. Evidence-based patient-centered interviewing. *JCOM-WAYNE PA*. 2001;8(7):28-34.
26. Duke P, Frankel RM, Reis S. How to integrate the electronic health record and patient-centered communication into the medical visit: a skills-based approach. *Teach Learn Med*. 2013;25(4):358-365. doi:10.1080/10401334.2013.827981
27. Frankel RM, Stein T. Getting the most out of the clinical encounter: the four habits model. *J Med Pract Manage*. 2001;16(4):184-91.
28. Braddock CH, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice. *JAMA*.1999;282(24):2313-2320. doi:10.1001/jama.282.24.2313
29. Canavan C, Holtman MC, Richmond M, Katsufrakis PJ. The quality of written comments on professional behaviors in a developmental multisource feedback program. *Acad Med*. 2010;85(10 suppl):106-109. doi:10.1097/ACM.0b013e3181ed4cdb
30. Carpenter SK, Cepeda NJ, Rohrer D, Kang SH, Pashler H. Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. *Educ Psychol Rev*. 2012;24:369-378. doi:10.1007/s10648-012-9205-z
31. Nowell LS, Norris JM, While DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. *Int J Qual Methods*. 2017;16:1-13. doi:10.1177/1609406917733847
32. The Coalition for Physician Accountability. *The Coalition for Physician Accountability's Undergraduate Medical Education-Graduate Medical Education Review Committee (UGRC): Recommendations for Comprehensive Improvement of the UME-GME Transition*. Accessed March 13, 2022. https://physicianaccountability.org/wp-content/uploads/2021/08/UGRC-Coalition-Report-FINAL.pdf
33. McConville JF, Woodruff JN. A shared evaluation platform for medical training. *N Engl J Med*. 2021;384(6):491-493. doi:10.1056/NEJMp2031317
34. AACOM, AAMC, ACGME, ECFMG/FAIMER. *Transition in a Time of Disruption. Practical Guidance to Support the Move from Undergraduate Medical Education to Graduate Medical Education*. Published March 2021. Accessed December 13, 2022. https://www.aamc.org/media/51991/download
35. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. *Acad Emerg Med*. 2008;15(11):988-994. doi:10.1111/j.1553-2712.2008.00227.x
36. Young JQ, Sugarman R, Schwartz J, O'Sullivan PS. Faculty and resident engagement with a workplace-based assessment tool: use of implementation science to explore enablers and barriers. *Acad Med*. 2020;95(12):1937-1944. doi:10.1097/ACM.0000000000003543
37. Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Adv Health Sci Educ Theory Pract*. 2016;21(2):455-473. doi:10.1007/s10459-015-9614-0
38. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Med Educ*. 2009;43(10):989-992. doi:10.1111/j.1365-2923.2009.03438.x
39. Hemmer PA, Dadekian GA, Terndrup C, et al. Regular formal evaluation sessions are effective as frame-of-reference training for faculty evaluators of clerkship medical students. *J Gen Intern Med*. 2015;30(9):1313-1318. doi:10.1007/s11606-015-3294-6

Jennifer R. Kogan, MD, is Associate Dean, Student Success and Professional Development, and Professor of Medicine, Perelman School of Medicine, University of Pennsylvania; **Lisa N. Conforti, MPH,** is Research Associate for Milestones Evaluation, Accreditation Council for Graduate Medical Education (ACGME); and **Eric S. Holmboe, MD,** is Chief Research, Milestone Development, and Evaluation Officer, ACGME.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank the following individuals: Denise LaMarra, MS, CHSE, Janice Radway, BA, Marc Shalaby, MD, and the Perelman School of Medicine Standardized Patient Program, including the actors for assistance developing stimulus videos; Carol Chou, MD, Nicole Defenbaugh, PhD, Richard Frankel, PhD, Benjamin Kinnear, MD, MEd, Denise LaMarra, MS, CHSE, and Leigh Simmons, MD, for assistance developing the stimulus videos; Stephanie Taitano and the Perelman School of Medicine Faculty Affairs and Professional Development Staff for their assistance creating the spaced learning modules; Anthony R. Artino Jr, PhD, for his help with the end of study survey; and focus group leaders Elizabeth Bernabeo, MPH, Justin Bittner, Laura Hirshfield, PhD, and Judy Shea, PhD.

Corresponding author: Jennifer R. Kogan, MD, Perelman Center for Advanced Medicine, koganj@pennmedicine.upenn.edu

Received April 6, 2022; revisions received June 27, 2022, and December 3, 2022; accepted December 6, 2022.