How Teachers Adapt Their Cognitive Strategies When Using Entrustment Scales

Milena Forte, MD, Natalie Morson, MD, MSCCH, Natasha Mirchandani, MD, Batya Grundland, MD, MEd, Oshan Fernando, MA, PhD, and Warren Rubenstein, MD

Abstract

Purpose
Rater cognition is a field of study focused on individual cognitive processes used by medical teachers when completing assessments. Much has been written about the role of rater cognition in the use of traditional assessment scales. Entrustment scales (ES) are becoming the tool of choice for workplace-based assessments. It is not known how moving to an entrustment framework may cause teachers to adapt their cognitive rating strategies. This study aimed to explore this gap by asking teachers to describe their thinking when making rating decisions using a validated ES.

Method
Using purposive sampling, family medicine teachers supervising obstetrical care were invited to participate in cognitive interviews. Teachers were interviewed between December 2018 and March 2019 using retrospective verbal protocol analysis. They were asked to describe their experiences of rating residents in the last 6 months using new ES. Constructivist grounded theory guided data collection and analysis. Interviews were recorded, transcribed, and analyzed iteratively. A constant comparative approach was used to code and analyze the data until consensus was reached regarding emerging themes.

Results
There was variability in how teachers used the ES. Faculty describe several ways in which they ultimately navigated the tool to say what they wanted to say. Four key themes emerged: (1) teachers interpreted the anchors differently based on their cognitive framework, (2) teachers differed in how they were able to cognitively shift away from traditional rating scales, (3) teachers struggled to limit assessments to a report on observed behavior, and (4) teachers contextualized their ratings.

Conclusions
Variability in teachers’ interpretation of learner performance persists in entrustment frameworks. Rater’s individual cognitive strategies and how they observe, process, and integrate their thoughts into assessments form part of a rich picture of learner progress. These insights can be harnessed to contribute to decisions regarding readiness for unsupervised practice.

Although competency-based medical education (CBME) strives to provide an overall objective assessment of progress, it largely relies on teachers to provide judgments of learners, which are inherently subjective.1–3 Assessor ratings are known to be variable even when based on the same observation of performance.4,5 Variability does not necessarily represent bias or rater error. It may also reflect the unique ways in which raters interpret the task or judge its completion.6 Idiosyncrasy can be meaningful, as workplace-based assessment is context dependent. Previous research suggests that even when provided with specific direction, clinical supervisors use their subjective internal cognitive strategies when assessing trainees.7,8 Rater cognition is a field of study which has evolved to focus on the individual cognitive processes used by medical teachers when completing assessments.7 Much has been written about the role of rater cognition in the use of traditional assessment scales.7,10 Gauthier et al conducted a literature review and subsequently proposed a framework within which all conceptualizations of rater (or assessor) cognition to date could be situated.9 Their framework included 3 phases: observation, processing, and integration. Observation includes how teachers generate impressions about the learner and also how they choose what aspect of their performance they will focus their assessments on. Yeates et al propose that “differential salience” may lead to differing interpretations of a similar performance, referring to the fact that different teachers may attribute variable importance to different aspects of the task they are observing (e.g., technical knowledge vs communication with the patient).9 This reflects what the teacher values as well as what they have observed. Processing is the phase where teachers attempt to make meaning of what they have observed by comparing it with a reference or standard when deciding how to rate them. Finally, integration occurs when teachers translate their judgment into a form or scale.10 We suggest that in addition to rater cognition, rater emotion may also play a role in influencing assessments and serve as a legitimate source of information.11

Entrustment scales (ES) are becoming the tool of choice for workplace-based assessments in many CBME-training programs. ES attempt to reframe assessments from a judgment of competency to a question of whether a learner is ready to perform a specific task at a defined level of supervision in the journey toward unsupervised practice.12–14 Several factors influence how teachers grant clinical trust in the workplace.15–22 These include the nature of the task, the context of the clinical environment, as well as trainee behavior. Teacher attitudes, expertise, and familiarity with the learner.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Milena Forte, 60 Murray St., 4th Floor, Box 25, Toronto, ON M5T 3L9, Canada; telephone: (416) 586-4800, ext. 3190; email: milena.forte@sinahealth.ca; Twitter: @MFischia.
Entrustment and Supervision
can also impact entrustment decisions. However, it is not known how moving to an entrustment framework may cause teachers to adapt their cognitive rating strategies. We aimed to explore this gap by asking teachers to describe their thinking when making rating decisions using a validated ES.

Method
This study was approved by the institutional ethics review board of Mount Sinai Hospital (MSH) on June 8, 2018. Using a constructivist grounded theory approach, we aimed to provide a conceptualization of the variable ways in which teachers’ cognitive strategies underpinned how they formulated and recorded assessments of trainees. Through cognitive interviews and retrospective verbal protocol analysis, we explored how faculty used the new ES to assess learners. Specifically, we used a “think aloud” protocol to explore what the anchors meant to them and how they decided when to use them. This allowed us to gain an understanding of the cognitive processes teachers employed when making rating decisions by asking them to describe rather than explain their thoughts. We also asked the teachers about how their cognitive strategies differed when using the ES as compared with traditional assessment scales.

We adapted a previously validated ES to an online tool used to assess trainees in a family medicine (FM) maternity care setting. Our tool required faculty to choose from 5 anchors that described residents’ level of independence for the specified tasks ranging from “I had to do it” to “I did not need to be there.” The anchors were listed vertically, without numbers. The ES differed from the prior assessment instrument, which was a traditional scale oriented from left to right with numbered anchors ranging from unsatisfactory to excellent (see Figure 1).

Participants and setting
MSH houses the labor and delivery unit (LDU) for 3 teaching hospitals within the University of Toronto. During the study period, 75 FM residents from all 3 sites completed their intrapartum maternity care rotation at MSH where they were primarily supervised by FM staff. Each resident also learns about antenatal care at their primary FM training site. In July 2018, all faculty supervising residents providing antenatal or intrapartum care at these sites were introduced to the new ES via a brief presentation at a meeting and follow-up emails.

Figure 1 Domains and anchors for Family Medicine Maternity Care Assessment Tools.
completed representing faculty from all 3 sites, representing a range of backgrounds and years of experience (see Table 1).

### Analysis

We used a constant comparative approach to analyze the data. The team met after the first 2 interviews to review the transcripts, identify early themes, and modify the interview guide. Transcripts were reviewed for new themes by the team as interviews progressed, and an early coding framework was developed. Having read all transcripts at the end of the data collection phase, the team met to perform line-by-line coding of 2 transcripts, which helped further refine the coding scheme. The RA then coded all of the transcripts using NVivo 11 (QSR International, Melbourne, Australia). Monthly team meetings over a period of 9 months followed to discuss emerging findings until no new themes arose.

### Results

There was a tension between the objective structure of the ES and the subjectivity inherent in teacher assessments.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participant Demographic Characteristics</strong></td>
</tr>
<tr>
<td><strong>Demographic characteristics</strong></td>
</tr>
<tr>
<td><strong>Sex, no. (%)</strong></td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td><strong>Practice site</strong></td>
</tr>
<tr>
<td>MSH</td>
</tr>
<tr>
<td>TWH</td>
</tr>
<tr>
<td>TWHFHT-GC</td>
</tr>
<tr>
<td>WCH</td>
</tr>
<tr>
<td><strong>Years in practice</strong></td>
</tr>
<tr>
<td>≤ 10</td>
</tr>
<tr>
<td>11–30</td>
</tr>
<tr>
<td>&gt; 30</td>
</tr>
<tr>
<td><strong>Deliveries per year</strong></td>
</tr>
<tr>
<td>≤ 10</td>
</tr>
<tr>
<td>11–60</td>
</tr>
<tr>
<td>&gt; 60</td>
</tr>
<tr>
<td>N/A (antenatal)</td>
</tr>
</tbody>
</table>

Abbreviations: MSH, Mount Sinai Hospital; TWH, Toronto Western Hospital; TWHFHT-GC, Toronto Western Hospital Family Health Team-Garrison Creek; WCH, Women’s College Hospital; N/A, not applicable.

In other cases, teachers compared the resident’s performance with what their own would have been at that level, or to what was expected of them at the same stage of training. Some described rating residents’ performances relative to other residents they had worked with, particularly if they saw themselves as experienced teachers: “I think just because of all the years I’ve worked there I can understand, usually after a little while where the resident is compared to their peers” (T3).

Teachers often rated residents based on how they felt they should perform for the stage of training they were at: “I use anchors in mine: early postgraduate year (PGY)-1, late PGY-1 or early PGY-2, late PGY-2….” (T13).

Some also described how they looked for improvement over time, comparing a resident they had worked with previously with their last performance and looking for growth:

> … If its someone that I’ve worked with before I’m looking to see whether or not I feel that they are progressing along a certain trajectory, so have they improved compared to the last time I worked with them. (T7)

Finally, for some teachers, the only way they would choose “I did not need to be there” was if they felt the resident was ready for unsupervised practice: “If I was going to give someone ‘I did not need to be there’ … then they would need to able to function as an independent physician” (T3).

### Theme 2: Teachers differed in how they were able to cognitively shift away from traditional rating scales

For decades, our faculty have used assessment tools that rank learners (e.g., unsatisfactory to excellent, or, not competent to proficient). The move away from these traditional assessment scales to an entrustment framework required a significant cognitive shift. This “retraining the evaluation brain” came more easily to some teachers than to others. The teachers who made the shift seamlessly described how the anchors were better aligned with how they preferred to communicate their assessment of (and to) the resident: “It is quite a bit better match with what is going on in my head” (T10). Some participants noted that the ES freed them to use a wider range of ratings. In particular, it allowed them to use the lower end of the scale without fear of upsetting the learner as it felt less value laden than traditional scales. Teachers also commented that the entrustment anchors allowed them to be more honest with their feedback since it was a reflection of a shared experience with the learner and an “observation of their skill level” rather than a “judgment on them” (T3).

> … if I have to put “I had to do it” or “I had to talk them through it” it wouldn’t be a surprise … and I don’t think they would be offended by that, I hope. (T3)
But for some, the shift proved challenging and the ES did not enhance their ability to say what they wanted to say:

I don't understand the descriptors very much. Whereas the old anchors were poor, fair, average, above average, excellent, or something like that, then I would say he's at average for a first year…. (T8)

Teachers took liberties with the scale when they felt the anchors did not allow them to express what they wanted to convey. They described making a *mental translation* of the anchors and ultimately reverting to the more familiar 1–5 Likert-type scales in their minds: “If I didn’t feel it totally fit the situation, I would sort of just use it as a 5-point scale, 5 meaning excellent and 1 being like bad” (T3).

Despite the fact that the scale was visually presented as a vertical drop-down menu with no numeric scale attached, many teachers rearranged it in their minds to fit with a “left to right” or “weak to strong” orientation. It was hard for these teachers to make the cognitive shift away from the left hand of the scale representing the “low end of the scale.” Therefore, they avoided choosing “I had to do it.”

**Theme 3: Teachers struggled to limit their assessments to a report on observed behavior**

Many teachers struggled to limit their assessments to a report on observed behavior (despite being directed to do so), choosing instead to use the form to provide summative assessments. In cases when faculty felt the resident performance did not fit their global impression of them, they hesitated to use the anchors to simply report on the observed activity. For instance, when the resident did not perform a task independently, some teachers expressed concern that choosing “I had to do it myself” might contribute to a poor overall grade when they actually felt the resident was not yet ready to always perform this task on their own, they were uncomfortable choosing “I did not need to be there,” feeling that this implied the student could now be entrusted to complete this task independently moving forward.

Teachers often described rating according to how they thought a resident might perform in some future situation as opposed to reporting on what they observed the resident do:

When I’m filling out these evaluations I’m using a combination of the interactions that we had, the kind of case we had, and then also, sort of forecasting even in the future would I just never need to be there? (T12)

Participants often described discomfort choosing “I did not need to be there” because they anticipated that they may need to be there in the future. They wondered, “Yes but, what if?” and opted to choose “I needed to be there just in case” to signal that the resident was not ready to handle more complex situations.

**Theme 4: Teachers needed to contextualize ratings**

Many teachers described that the ES anchors did not allow them to tell the whole story: “… I wrote actually something at the end because these sorts of options didn’t quite capture what I would have wanted to say if I could say it” (T3).

Several felt the need to use the comments section to qualify their evaluation and provide some explanation to accompany their ratings because the ES either did not fully capture the essence of what they wanted to say or was at risk of being misinterpreted: “It felt very hollow without it (comments) … I mean, it would have been completely out of context” (T5).

A learner who was excellent overall may receive a rating of “I had to do it” for a case she had never encountered before. Conversely, another resident may be rated the same in a situation that they should know how to manage independently. In one case, the chosen anchor is completely appropriate and expected, while in the other it may indicate a deficiency. Teachers also indicated that the context of the clinical scenario needed to be described and considered, particularly if it was a challenging case and they chose an anchor such as “I had to talk them through it.” The need for comments was particularly evident when assessing professional attitudes such as collaborator or health advocate. Teachers felt these needed more descriptive context and many struggled to fit their assessments onto the scale:

… For the procedural skills, it’s very specific and very, I think, black and white. Like, either they could do it or they couldn’t. Whereas with the professional attitudes it’s more, you know, subjective…. (T14)

**Discussion**

Teacher factors impact entrustment decisions. Ours is the first study to our knowledge that systematically explores the cognitive framework that teachers employ when using ES in the clinical workplace. Previous studies were based on traditional assessment scales and largely examined how teachers rated a scripted performance they observed on video, or standardized encounters under highly controlled parameters.9, 10 Our study reports on workplace-based encounters and reveals numerous ways that teacher subjectivity is inherent in the assessment process, particularly in the *processing* and *integration* phases of Gauthier’s framework.9 Previous studies looking at the role of rater cognition detail how teachers use a variety of frames of reference when rating learners using traditional assessment scales.9,10 Our participants vividly described how they continued to use their own inherent cognitive frameworks to *process* or benchmark resident performance despite using an ES that asked them to simply report on the level of supervision required for a specific task. Although our ES asked for a seemingly objective recount of learner performance, teachers ultimately remained bound to the same subjectivity they bring to other rating scales.

All assessment tools have a structure. Regardless of the assessment framework being used, there can exist a tension between this structure and the nuances teachers wish to capture and document. While the tool appears to facilitate an objective rating, an evaluation is still ultimately a subjective account of the learner’s ability based on both the assessor’s perception of the learner’s skill and a reflection of the teacher’s values and cognitive framework. The switch to
an ES from traditional rating scales did not change that tension. It only changed the structure within which teachers could integrate their assessments.

Viewed from a social constructivism perspective, the subjective opinion of the teacher is layered onto the structure of the assessment tool to create a co-constructed version of their level of progress. The tool defines the parameters to be assessed, and the teacher defines the content that is reported or judged. The ES better aligned with some teachers’ translations of their assessments and required a pronounced cognitive shift for others.

In addition to maneuvering around the ES to say what they wanted to say, teachers spoke about the need to contextualize their assessments using the comments field. This has been reported elsewhere. Interestingly, in our study, this was particularly true when the anchors would have led them to rate a learner on the “low” end of the scale. The teachers were not always able to disentangle the notion of the “left hand of the scale” indicating a poor rating. This reflects a persistence in seeing the ES as a judgment of the residents’ overall competence rather than a statement on what they were able to accomplish on their own in a particular clinical encounter.

Teachers in our study often struggled to limit their assessments to reports on observed behaviors, or what ten Cate calls “ad hoc entrustment decisions,” choosing instead to use the entrustment anchors to provide a summative or global assessment of a resident’s competency. This echoes work that found ratings using ES reflected an “averaging of all prior experiences” as opposed to performance on a specific task. This evaluative imperative appears to hold true regardless of whether teachers are instructed to frame the assessment as formative or summative. Despite using a retrospective ES, our teachers tempered their ratings of what they had observed by asking themselves how a resident might perform in the future, not how they performed in the past. We must therefore not assume that ad hoc entrustment decisions routinely made by teachers in the workplace always correlate with the report on observed behavior teachers submit after the fact. The latter is an intentional rating and represents information the teacher wishes to convey about the learner, knowing it will form part of that learner’s evaluation record. The former is a reflection of what the teacher felt comfortable letting the learner do, based on a combination of the learner’s readiness to perform it and the teacher’s readiness to let them perform it.

We have detailed the multitude of ways that teachers variably employed cognitive strategies when using ES for workplace-based assessments. Our data suggest that a deeper understanding of how faculty interpret assessment tools including ES is vital to educators and programs, perhaps more vital than efforts aimed at refining anchors or constructs used in a given tool. Past efforts at faculty development around assessment have often focused on either training teachers as to what individual anchors mean or on developing shared mental models. These efforts to help harmonize teachers’ understanding of the tools so that they may function more objectively have by and large shown very modest effects. Ultimately, it is the clinical teachers who guide programs in their decisions regarding learner autonomy on the journey toward unsupervised practice. We propose that any faculty development on ES must address the diversity of teacher perspectives and acknowledge the inherent subjectivity they bring to assessments via their cognitive frameworks. Making teachers aware that they may be using one (of many) frameworks to benchmark performance may allow them to better situate their feedback and perhaps guide their formative feedback conversations with learners. Faculty that collate evaluations for summative purposes might be encouraged to consider the breadth of frameworks among their teachers and how these may affect individual assessments. In short, rather than aim to eliminate teacher subjectivity, programs could aim to harness it to capture a rich picture of trainee progress in a multitude of contexts.

**Strengths and limitations**

By providing faculty with a recent evaluation they completed and asking them to think aloud about their choices, we were able to capture the multiple ways in which they interpreted the anchors. This allowed us to capture a closer approximation of their real-time thinking than a simulated trigger case would have.

Our sample size was relatively small and drawn from one institution. This is a common challenge in rater cognition research. However, saturation of themes was reached and it represented a diverse group of teachers from 3 academic teaching units as well as community preceptors. Given that our teachers were exclusively family physicians, further work in this area would ideally sample perspectives from teachers in a variety of disciplines and other institutions. Future directions could be aimed at exploring the experience of being assessed with ES from the resident perspective.

**Conclusions**

We systematically explored the cognitive strategies that teachers employed when using ES in a 2-part interview process. We conducted cognitive interviews and also provided an informal opportunity for teachers to reflect on their own assessments.

Our study found that variability in rater cognition seems to persist when using ES. While the addition of ES to the assessment armamentarium holds much promise, teachers’ individual cognitive rating strategies, including how they observe, process, and integrate learner performance when completing assessments, must be considered. ES require teachers to make a cognitive shift from traditional assessment forms. The need for teachers to be able to properly contextualize their assessment of learners remains a key feature of any assessment tool, including those using an entrustment framework. The rich variability in assessments provided by teachers should contribute to a learner’s path to unsupervised practice.

**Funding/Support:** Funding for this study was provided by the Sinai Health System, Department of Family Medicine Research and Education Fund.

**Strengths and limitations**

Strengths and limitations By providing faculty with a recent evaluation they completed and asking them to think aloud about their choices, we were able to capture the multiple ways in which they interpreted the anchors. This allowed us to capture a closer approximation of their real-time thinking than a simulated trigger case would have. Our sample size was relatively small and drawn from one institution. This is a common challenge in rater cognition research. However, saturation of themes was reached and it represented a diverse group of teachers from 3 academic teaching units as well as community preceptors. Given that our teachers were exclusively family physicians, further work in this area would ideally sample perspectives from teachers in a variety of disciplines and other institutions. Future directions could be aimed at exploring the experience of being assessed with ES from the resident perspective.

**Conclusions**

We systematically explored the cognitive strategies that teachers employed when using ES in a 2-part interview process. We conducted cognitive interviews and also provided an informal opportunity for teachers to reflect on their own assessments.

Our study found that variability in rater cognition seems to persist when using ES. While the addition of ES to the assessment armamentarium holds much promise, teachers’ individual cognitive rating strategies, including how they observe, process, and integrate learner performance when completing assessments, must be considered. ES require teachers to make a cognitive shift from traditional assessment forms. The need for teachers to be able to properly contextualize their assessment of learners remains a key feature of any assessment tool, including those using an entrustment framework. The rich variability in assessments provided by teachers should contribute to a learner’s path to unsupervised practice.

**Funding/Support:** Funding for this study was provided by the Sinai Health System, Department of Family Medicine Research and Education Fund.
Other disclosures: None reported.

Ethical approval: This study was approved by the institutional ethics review board of Mount Sinai Hospital on June 8, 2018. REB approval #18-0081-E.

Previous presentations: The abstract of an earlier version of this article was presented at the Canadian Conference on Medical Education (CAME) in April 2020.

M. Forte is assistant professor, Postgraduate lead, Office of Education Scholarship, Maternal Care lead, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; ORCID: https://orcid.org/0000-0001-5387-0184.

N. Morson is lecturer, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

N. Mirchandani is assistant professor, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

B. Grundland is assistant professor and associate postgraduate program director, Curriculum & Remediation, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

O. Fernando is research associate, Office of Education Scholarship, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

W. Rubenstein is associate professor, Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

References


