Advancing Our Understanding of Narrative Comments Generated by Direct Observation Tools: Lessons From the Psychopharmacotherapy-Structured Clinical Observation

John Q. Young, MD, MPP, PhD
Rebekah Sugarman, AB
Eric Holmboe, MD
Patricia S. O'Sullivan, EdD

ABSTRACT

Background While prior research has focused on the validity of quantitative ratings generated by direct observation tools, much less is known about the written comments.

Objective This study examines the quality of written comments and their relationship with checklist scores generated by a direct observation tool, the Psychopharmacotherapy-Structured Clinical Observation (P-SCO).

Methods From 2008 to 2012, faculty in a postgraduate year 3 psychiatry outpatient clinic completed 601 P-SCOs. Twenty-five percent were randomly selected from each year; the sample included 8 faculty and 57 residents. To assess quality, comments were coded for valence (reinforcing or corrective), behavioral specificity, and content. To assess the relationship between comments and scores, the authors calculated the correlation between comment and checklist score valence and examined the degree to which comments and checklist scores addressed the same content.

Results Ninety-one percent of the comments were behaviorally specific. Sixty percent were reinforcing, and 40% were corrective. Eight themes were identified, including 2 constructs not adequately represented by the checklist. Comment and checklist score valence was moderately correlated (Spearman's rho = 0.57, $P < .001$). Sixty-seven percent of high and low checklist scores were associated with a comment of the same valence and content. Only 50% of overall comments were associated with a checklist score of the same valence and content.

Conclusions A direct observation tool such as the P-SCO can generate high-quality written comments. Narrative comments both explain checklist scores and convey unique content. Thematic coding of comments can improve the content validity of a checklist.

Introduction

The adoption of competency-based frameworks in medical education has highlighted the need for workplace-based assessment with a dual focus on the assessment of learning (ie, summative feedback) and the assessment for learning (ie, formative feedback).[1–3] In this context, performance assessment based on direct observation of a trainee-patient encounter has become increasingly important. Direct observation tools have been developed for general clinical skills (eg, miniCEX) and for focused tasks, such as electromyography, teamwork, laparoscopy, ultrasound-guided anesthesia, handoffs, and follow-up visits.[4–10] These tools typically include either a list of behaviors (ie, checklist) or competencies that are rated on a quantitative scale as well as space for narrative comments.

While many direct observation tools include rating scales and space for narrative comments, validity arguments have largely focused on the quantitative scores generated by the rating scales.[4,11] Much less is known about the quality of the comments, a critical component of many direct observation tools. In fact, most of our knowledge about the quality of feedback generated by workplace-based assessments comes from research on end-rotation evaluations or multisource feedback (ie, not direct observation of a single clinical encounter).[12–14] These studies showed that narrative comments can provide helpful guidance to learners and enhance summative decisions, especially when combined with quantitative ratings.[15–17] Research has also identified challenges with narrative comments generated by end-rotation evaluations, including variable quality and the purposeful use of vague or coded language that can be difficult to interpret.[13,18]

DOI: http://dx.doi.org/10.4300/JGME-D-19-00207.1

Editor's Note: The online version of this article contains narrative comments and exemplar comments by theme from the Psychopharmacotherapy-Structured Clinical Observation.

Yet the findings on narrative comments from end-rotation and multi-source feedback may not apply to direct observation. Typically, these types of assessments use different tools and occur in different contexts. For example, the end-rotation and multi-source feedback types typically ask a rater to synthesize impressions (not necessarily based on direct observation of clinical encounters) gathered over weeks or months. In contrast, feedback based on direct observation occurs during or immediately after a single observed patient encounter. Only a few studies have examined qualitative feedback generated by direct observation. One study (mini-CEX) focused on verbal (not written) comments and found that faculty provided specific recommendations, but underutilized the feedback methods of self-assessment and action planning.[19] In a subsequent study, a modified mini-CEX generated comments that were specific.[20] In order to more fully appraise the evidence for the utility and validity of assessments generated by direct observation tools, it is important to determine the quality of the narrative feedback.[21]

In addition, we know very little about the relationship between the information provided by the rating scales and the comments. Two studies (end-rotation evaluations and end-of-shift) found an association between lower quantitative scores and the presence of more corrective comments.[22,23] A third study, in which faculty rated the video of an intern interviewing a standardized patient, found that the valence (ie, reinforcing versus corrective) of narrative comments showed moderate to strong correlations to quantitative scores.[24] But research has not yet examined the relationship between the quantitative scores and the narrative comments generated by direct observation tools in the workplace. To what extent is the information each conveys similar or different? Answering this question will improve our understanding of the role of written comments vis-à-vis the rating scales, including to what extent the comments expand on the scores versus contribute new content. Moreover, if the comments are "thick," they should cover the essential competencies and can be used to assess the content validity of the rating scale.[21]

We sought to address these 2 gaps in the literature—the quality of written comments generated by a direct observation tool and the relationship between the quantitative and narrative information. We used the comments generated by the Pharmacotherapy-Structured Clinical Observation (P-SCO), a direct observation tool in psychiatry, with evidence for validity.[25–27] This study has 3 aims: (1) Analyze the quality of the narrative comments generated by the P-SCO; (2) Characterize the themes most commonly captured by the narrative comments; and (3) Examine the relationship between the narrative comments and the checklist scores.

What was known and gap
Direct observation tools for workplace-based assessment often include narrative comments as well as quantitative scores, but research has focused on the quantitative aspects.

What is new
An assessment of the quality of the narrative comments and their relationship to the checklist scores of a direct observation tool used for psychiatry residents.

Limitations
Single institution and specialty limits generalizability.

Bottom line
A direct observation tool can generate high-quality written comments that both explain the checklist scores and add new content.

Methods

Design

This is a mixed-methods study that uses an already existing data set of 601 completed P-SCOs. The P-SCOs were completed by 11 faculty on 64 residents distributed over 4 academic years (2008–2012). While P-SCOs completed more recently were not collected for analysis, the P-SCO itself remains unchanged and is used in multiple training programs. The activity that the P-SCO assesses is considered a primary activity of a psychiatrist and has recently been identified as a core end-of-training entrustable professional activity.[28]

Setting

The P-SCO was implemented in the outpatient medication management clinics of a university-based psychiatric hospital. The clinics provided the primary and final required experience in ambulatory pharmacotherapy for all 16 or 17 (depending on the academic year) postgraduate year 3 (PGY-3) residents in the program. Each resident was assigned to a half-day clinic, including a 30-minute case conference and 3.5 hours of patient contact. Each clinic had 4 to 5 residents with 2 attending physicians. The attendings differed between the clinics. The resident-attending cohort remained intact for each resident's 12-month experience.

Intervention

Prior studies of the P-SCO have shown evidence for validity with respect to its content, internal structure, and association of its scores with resident experience.[26,29] The P-SCO had 27 checklist items that represent the essential tasks of a medication

management visit in psychiatry. Rather than a yes/no scale seen in some checklists, the P-SCO used a 4-point rating scale to capture the continuum of "done": 1, not done; 2, done with suggestions for improvement; 3, done well (meets expectations); and 4, done extraordinarily well (inspires me to do the same). The scale included a fifth option: N/A (not applicable). In addition, the P-SCO had a space for faculty to record narrative comments on strengths and areas for improvement.

During the pre-clinic case conference, faculty decided which trainees to observe. After the patient encounter, faculty completed the paper P-SCO (checklist and narrative comments), provided verbal feedback to the trainee, and returned the completed P-SCO to the clinic director's mailbox. The clinic director made a copy for administrative purposes and returned the original to the trainee. Faculty and residents received training in the use of P-SCO at the beginning of each academic year. The training included watching a videotape of a resident session, completing the P-SCO, comparing how each person scored the videotaped resident, exploring the purpose of the direct observations and feedback, and then reviewing expectations for the upcoming academic year. Each resident received on average 9 completed P-SCOs over any given academic year.

Procedures

The completed P-SCOs were deidentified by assigning a unique code to each faculty member, resident, and completed observation. We used a random number generator to sample 25% of the completed P-SCOs from each year for a total sample of 152. Two authors (J.Q.Y. and R.S.) independently coded the comments on each P-SCO. The unit of analysis was a discrete comment. We defined a comment as a grouping of words (eg, partial sentence, full sentence, or multiple sentences) focused on a unique concept or behavior. The first author was an attending physician in the clinic who performed P-SCOs.

In evaluating the written comments, we focused on 3 attributes that prior research has established as important dimensions of quality: specificity, valence, and content.[14,30] Coding options for specificity were specific, general, or indeterminate. A comment was designated specific when it described a behavior of the trainee with enough detail that the trainee could act on the information (eg, "When screening for adherence, you might ask 'How many missed doses?'"). A typical general comment was "great job on the interview." We also coded each comment for its valence or polarity (ie, negative or positive). Valence options were reinforcing (endorsement of the behavior), corrective, or indeterminate. For content, the authors developed an initial coding scheme together based on the behaviors (eg, assessing medication adherence or managing medication adverse effects) identified in the checklist. Two authors independently coded 2 to 5 P-SCOs at a time and then compared text deemed a discrete comment and the assigned codes (specificity, valence, and content) for each comment. Differences were resolved through consensus, and the content code book was modified iteratively. Modifications included the addition of a new code or the lumping or splitting of previous codes. By the 25th observation, the reviewers were no longer identifying new codes and were rarely disagreeing on code assignment. At that point, the 2 authors independently coded each completed P-SCO in batches of 25. Assigned codes were compared and differences resolved through consensus after each batch. At the end, the dataset was reanalyzed using the final coding scheme.

Data Analysis

Quality of the Comments Comments that had indeterminate valence and/or content were excluded. We calculated the proportion and mean number of comments per observation that were specific versus general and reinforcing versus corrective.

Primary Themes General comments did not, by definition, address a specific behavior and were excluded from the content analysis. The authors independently clustered the content codes into proposed themes. The authors compared themes and, through discussion and review of the discrete comments associated with each code, developed consensus on themes. We calculated the prevalence of each theme and the proportion of reinforcing and corrective comments within each theme.

Relationship of the Comments and the Checklist Scores In order to determine the extent to which the checklist scores (high or low) and the narrative comments (reinforcing or corrective) provide similar or different information, we ran 3 types of analyses. First, 2 authors (J.Q.Y. and R.S.) examined the extent to which each primary narrative theme represented constructs captured by the checklist.

Second, to assess whether the valence of the comments and checklist scores were aligned, we adapted a methodology used by Yeates et al.[24] For each P-SCO, we determined the overall valence of the comments by assigning scores of +1 for each

TABLE 1
Specificity and Valence of P-SCO Narrative Comments

Comment Specificity	Total		Reinforcing		Corrective	
	N (%)	Mean/Form (SD, Range)	N (%)	Mean/Form (SD, Range)	N (%)	Mean/Form (SD, Range)
Specific	697 (91)	4.6 (2.7, 0–13)	423 (86)	2.8 (1.9, 0–9)	274 (100)	1.8 (1.8, 0–9)
General	69 (9)	0.5 (0.7, 0–3)	69 (14)	0.5 (0.7, 0–3)	0 (0)	0 (0)
Total	766 (100)	5.1 (2.6, 0–13)	492 (100)	3.2 (2.0, 0–9)	274 (100)	1.8 (1.8, 0–9)

Note: There were a total of 779 comments from the sample of 152 completed observations. Thirteen comments were excluded from analysis because the content and/or valence (reinforcing or corrective) could not be interpreted.

reinforcing comment and -1 for each corrective comment and then calculated their sum. Similarly, we determined the valence of the checklist scores by assigning scores of +1 for each low checklist score (defined as a 1 or 2) and +1 for each high score (defined as 4) and calculated their sum. We excluded the score of 3 (done well–meets expectations) from this analysis because we did not have a narrative comment rating akin to "neutral." Because the checklist scale is ordinal, we performed Spearman rank order correlation.

Third, to further assess the alignment between the content and valence of the comments and the checklist scores, we calculated the proportion of low (defined as 1 or 2 out of 4) or high (defined as 4 out of 4) checklist scores that were accompanied by at least 1 corresponding corrective or reinforcing comment. Similarly, we calculated the proportion of corrective or reinforcing comments that were accompanied by at least 1 corresponding high or low checklist score. We performed this analysis at the level of the 3 factors or constructs that have been shown to underlie the P-SCO's 27-item checklist: affective tasks, cognitive tasks, and hard tasks.[29] Each checklist item was assigned to 1 of 3 factors as reported in a recently published study.[29] The authors used a process of consensus to assign each narrative code to 1 of the factors.

The Northwell Health Institutional Review Board deemed the study exempt from review.

We used Excel (Microsoft Corp, Seattle, WA) to calculate descriptive statistics and SPSS 24.0 (IBM Corp, Armonk, NY) for the correlational analyses.

Results

A total of 152 completed P-SCOs were randomly selected. The sample included 8 different faculty and 57 different residents. These P-SCOs yielded 779 comments. Thirteen comments had indeterminate valence and/or content and were excluded from analysis.

Quality

Six hundred and ninety-seven (91%) of the 766 narrative comments were behaviorally specific (TABLE 1). Each completed P-SCO yielded an average of 5.1 total comments and 4.6 (SD 2.7) specific comments, 2.8 (SD 1.9) specific reinforcing comments, and 1.8 (SD 1.8) specific corrective comments. Sixty-one percent (423 of 697) of the specific comments were reinforcing, while 39% (274 of 697) were corrective. All of the general comments were reinforcing.

Themes

Content analysis of the narrative comments identified 30 unique behaviors (ie, codes), and subsequent analysis yielded 8 primary themes (TABLE 2). Consensus on 5 themes emerged at the outset and paralleled the basic structure of a patient encounter: data gathering (obtains an interview history, elicits the narrative, builds rapport); assessment (assesses); and treatment (treats). "Educates" and "engages the patient" were separated into 2 themes because the comments for the former almost always described unidirectional information flow from the resident to the patient sometimes with teach-back, while the latter cluster captured bidirectional negotiations. The theme of "structures and manages the interview" emerged from the significant number of comments that focused on time management and transitions (how to begin and end).

The comments within each theme capture a set of specific behaviors that faculty thought important enough to warrant comment (TABLE 3; provided as online supplemental material). The ratio of reinforcing to corrective comments varied markedly by theme. The comments for "builds rapport" were overwhelmingly reinforcing and, to a lesser extent, so were the comments on eliciting the narrative and obtaining an interval history. The comments on "assesses the patient" and "engages the patient" were more often corrective. Very few comments focused on medical knowledge, clinical reasoning, or reviewing the chart.

Relationship Between Comments and Checklist Scores

Comparison of the narrative themes with the checklist items identified 2 themes not represented on the

TABLE 2
P-SCO Narrative Comments—Primary Themes (Most to Least Common)

Theme	Associated Codes	Reinforcing, N (%)	Corrective, N (%)	Total, N (%)	Ratio, Reinforcing to Corrective
Assesses	Assesses adherence, adverse effects, substance use, risk for violence and suicide, and response to treatment (including functional AA: status); employs symptom scales and mood charting; mental status examination; reviews charts; updates and modifies diagnosis as appropriate	52 (12)	95 (35)	147 (21)	0.5
Obtains an interval history	Obtains history (including target symptoms, medical or medication changes, intercurrent psychosocial stressors, progress in psychotherapy, collateral from family members)	70 (17)	32 (12)	102 (15)	2.2
Builds rapport	Treats patient with respect, establishes rapport (warm, empathic, caring), conveys hope, encourages ventilation of feelings regarding illness	90 (21)	9 (3)	99 (14)	10.0
Treats	Modifies treatment plan as necessary (including arranges for appropriate follow-up, attends to refills, manages adverse effects, addresses adherence problems), manages transitions in care, communicates with other members of the treatment team, medical knowledge, clinical reasoning	52 (12)	39 (14)	91 (13)	1.3
Educates	Educates the patient about diagnosis, prognosis, treatment, and/or adverse effects; educates the patient on self-care such as behavioral activation, exercise, sleep hygiene, coping skills, managing negative relationships	48 (11)	25 (9)	73 (11)	1.9
Elicits the narrative	Initial open-ended question, interviewing skills (including clarifying questions, appropriate use of open- and close-ended questions, follows the patients cues/affect, normalizes, links question to patient's affect)	53 (13)	14 (5)	67 (10)	3.8
Structures and manages the interview	Manages flow (sets the agenda, manages time, appropriate pacing and transitions, redirects the patient as necessary); stays in prescriber role	32 (8)	30 (11)	62 (9)	1.1
Engages	Engages patient in treatment planning (employs shared decision-making, attends to the patient's goals and values), solicits and addresses the patient's concerns, explores the patient's beliefs (confronts maladaptive beliefs, addresses the patient's ambivalence, utilizes motivational interviewing)	26 (6)	30 (11)	56 (8)	0.9

TABLE 3
P-SCO Narrative Comments—Exemplar Comments by Theme

Theme	Exemplar Corrective Comments	Exemplar Reinforcing Comments
Assesses	1. Suicidality—ask what mean by "not yet"—granted, patient said it in a light-hearted manner 2. Adherence: can ask "how many doses missed" rather than "have you missed" (normalize behavior)	1. Excellent how followed up on passive positive suicidal ideation that patient had expressed at last visit 2. Good how followed our patient's reference to having missed doses
Obtains an interval history	1. For sleep complaint, develop structured history: what time get in bed, how long until fall asleep, how many times awake and why, when out of bed in the AM, do you feel rested 2. Given that patient tags specific ongoing stressors, use some anticipatory guidance probing questions	1. Good combo of following patient's story, but also asking them to amplify 2. Liked your review of patient's challenges in life: marriage, work, anxiety
Builds rapport	1. Take more opportunity to follow up with questions about social issues (ie, new baby, work life) to build rapport 2. The sequence of sentences with pauses built tension and anxiety	1. Ability to remember details of patients' lives from session to session 2. Excellent balance in session of giving patient space and time to express emotions
Treats	1. We should be thinking about mobilizing other treatments for chronic psychosis 2. It is also reasonable to consider increasing/augmenting depression treatment given increased symptoms/impact and possibility that seizures will not be controlled for a while	1. Worked with other providers, great techniques of collaborative care 2. Excellent psychopharmacology changes/interventions
Educates	1. Be ready to give your recommendation, especially for indecisive patients 2. If you make a plan to possibly start a new medication, discuss in detail during session	1. Good education on his interest in "a shot to take away cravings" and the access that treating clinicians have to experimental treatments (ie, not much) 2. Clear and accurate explanations to patient in lay language
Elicits the narrative	1. "So how's your mood been since I saw you last," versus 3–4 sentences to get to question; work on more direct, simple question sentence 2. Sometimes/often a single blanket query can be useful to start	1. Tolerated conflict/tension in the room well 2. Good open ended question to start: "How's it been going?"
Structures and manages the interview	1. With patient as circumstantial and slightly pressured as this, should focus them more on symptoms 2. This patient can extend discussion. With them you have to say, "We need to end now" and then walk out	1. Appropriately focused for brief appointment (patient was 20 minutes late) 2. Extremely well-done job of maintaining the frame with a patient who challenges time and directed clinical assessment
Engages	1. Helpful to know what factors they believe would increase their risk of suicidality 2. Toward the end of session you might ask patient if there is anything else to talk about or have questions	1. I really liked your offer of choices 2. Good and detailed exploration of treatment plan currently in place and obstacles to better plan

checklist: "structures and manages the interview" and "engages the patient" (as distinct from building rapport and educating the patient). In addition, for the theme "eliciting the narrative," the comments identified behaviors both present (eg, starts with an open-ended question) and not present (eg, follows patient's cues) on the checklist, suggesting that this theme may be inadequately captured.

The valence of the narrative comments were significantly correlated with the valence of the checklist scores (Spearman's rho = 0.57, $P < .001$). In other words, completed P-SCOs with more

TABLE 4
Congruence Between Narrative Comments and Checklist Scores[a]

≥ 1 Narrative Comment Recorded (Reinforcing or Corrective)	≥ 1 Checklist Item Marked (High or Low)	
	Yes	No
Yes	200	195
No	88	429
Proportion of high or low checklist items accompanied by a corresponding narrative comment	69.4%	
Proportion of reinforcing or corrective narrative comments accompanied by a corresponding checklist score	50.6%	

[a] Six calculations were performed on each P-SCO: for each of the 3 factors, the number of low checklist scores and corresponding corrective comments and the number of high checklist scores and the number of corresponding reinforcing comments were calculated. For each of the 6 calculations, it was determined whether one or more high or low checklist scores were accompanied by one or more corresponding comments and vice-versa.

corrective comments in a given category had more low scores on the relevant checklist items, and P-SCOs with more reinforcing comments had more high scores on the relevant checklist items.

Finally, TABLE 4 shows the relationship between the content of the narrative comments and checklist items scored as high or low. Sixty-nine percent (200 of 288) of high or low checklist scores were accompanied by at least 1 corresponding narrative comment of the same valence. Narrative comments provided the resident with additional explanation and guidance for most high or low checklist scores. For example, a low checklist score for assessing adherence was accompanied by the comment: "Can ask, 'How many doses missed?' rather than 'Have you missed?' (normalize behavior)." On the other hand, only 51% (200 of 395) of corrective or reinforcing narrative comments were accompanied by a corresponding checklist score of the same valence. For a given observation, the narrative feedback addressed performance dimensions that the checklist scores did not; for example, a comment on engaging the patient in treatment planning might occur without a low checklist score. These findings suggest that the narrative and quantitative information provide overlapping but unique kinds of information.

Discussion

In this study, the P-SCO generated written comments that were behaviorally specific and clinically relevant, which are attributes associated with effective written feedback.[29] In addition, the comments included both reinforcing and corrective feedback, with more reinforcing than corrective, which represents an additional component of effective instruction.[30]

While we do not know the optimal ratio of reinforcing to corrective comments, the "magic" ratio of 5:1 is often cited in primary and secondary education (ie, a student needs to receive 5 affirmations to overcome the effects of a comment experienced as critical).[30] The ratio in this study was much smaller (3:2). However, the ideal ratio has not been empirically established for medical trainees, and "corrective" feedback delivered in a supportive way, as we hope is the case with the P-SCO, may be different than "critical" feedback in these other studies, which often are defined as "harsh" or "embarrassing."[30] Moreover, the quantity of comments per completed observation was high, especially when compared to prior studies of typical end-rotation assessments.[26] Taken together, these findings indicate that direct observation tools such as the P-SCO can generate narrative comments that support both assessment for learning and assessment of learning.

These findings contrast with studies of end-rotation evaluations, which tend to generate comments that are more commonly vague, both in content and valence.[13,18] This difference is likely multidetermined. First, the difference may derive from the nature of the task. Narrative comments based on direct observation are typically recorded soon after the event when the details are fresh, while end-rotation evaluations may be completed at a time more distant to the event and may not be based on direct observation of clinical encounters. Second, the difference may be related to the purpose of the assessment. Feedback from direct observation is most often formative and based on only a single encounter whereas end-rotation evaluations are more often summative. Faculty may perceive the former as lower stakes and psychologically easier to provide specific and corrective feedback. Third, the instrument itself, with specific behaviors embedded in the checklist, may help faculty structure their observation and generate specific feedback. Fourth, these P-SCOs were completed in the context of a yearlong faculty-resident longitudinal relationship, which may facilitate safety in the relationship and as a result more directness. Finally, the quality of the comments may relate to the culture of assessment that developed in

this clinic. The faculty participated in the initial development of the P-SCO, received annual training, and were given feedback on the quantity and quality of their completed P-SCOs from the clinic director. These factors may have facilitated greater faculty investment in the P-SCO process.

The themes captured by the comments are diverse and address many of the competencies related to pharmacotherapy.[31] The findings related to these themes have 2 important implications for other direct observation tools. First, thematic analysis of the narrative comments can help identify weaknesses in the checklist. For example, 2 of the themes—"engages the patient" and "structures the interview"—are not adequately captured by the P-SCO's checklist. While evidence for the content validity of the P-SCO exists, this suggests 2 possible additions to be considered in future iterations and more generally indicates the value that thematic analysis of comments from in vivo use can have in assessing the content validity of a checklist. In addition, the narrative comments did not address certain checklist items (eg, chart review, documentation, and communication with other members of the treatment team). We suspect this has to do with the context for observation (ie, the interview of a patient), which does not permit demonstration of these skills, since the associated activities occurred either before or after the patient encounter. These items may be safely removed from the checklist and assessed in a different context. These findings can be used to narrow the focus and enhance the content validity of the P-SCO. This kind of analysis of the narrative comments may be helpful to the development of other workplace-based assessment tools.

Second, thematic analysis of comments from direct observation tools may have an important role to play in identifying gaps in the curriculum. For example, themes for which there are relatively more corrective comments may indicate those competencies that residents have the most difficulty mastering. In this study, the 2 themes with more corrective than reinforcing comments were "assesses the patient" and "engages the patient." "Assesses" includes tasks such as utilizing symptom scales and addressing adherence, adverse effects, substance use, and suicide risk. This result is consistent with a prior study of the P-SCO in which these tasks were rated low on the checklist 30% of the time even at the end of the required outpatient training.[27] In general, programs of any specialty may use thematic coding of comments generated by their direct observation assessments to assess their own curriculum.

The quantitative analyses indicate moderate correlation between the valence of the narrative comments and the checklist scores, which provides further evidence for the validity of the P-SCO. Assessment of the alignment between the content of the comments and checklist scores suggests that each provides the learner with some unique information, a finding that was surprising. More than two-thirds of the time, narrative comments elaborated on high or low checklist scores, which likely made them easier for the learner to understand and act on. In addition, the narrative comments often provided feedback on behaviors not addressed by the checklist. The value of the narrative comments seems clear. What is less clear is the importance of the checklist. Completing the checklist comes at a cost, both in terms of time and rater cognitive load.[32] Learners may find comments more helpful than quantitative scores, and recent studies suggested that narrative comments can support summative judgements.[17,33] On the other hand, the checklist may prime the rater and thereby facilitate both a shared mental model between raters and the provision of behaviorally specific feedback. In addition, the completion of the checklist may lead to improvements in the raters' (faculty) own practice.[34] If the frame of reference provided by the checklist does result in higher-quality comments, then several important questions must be answered. Is provision of the checklist (rather than completion) adequate or is completion essential? If completion is essential, is there a threshold number of completed observations after which the checklist is sufficiently internalized by the faculty member and no longer necessary? As we look to sustainability of our direct observation assessment programs, these questions will be important to answer through future research on the P-SCO and other direct observation tools.

This study has limitations. The generalizability of the findings are limited by the use of residents and faculty at a single institution and specialty. Faculty without past experience in the P-SCO or without annual training likely perform differently. As one author who did the thematic coding was also a faculty rater, bias in coding decisions may have occurred, despite use of a second author for thematic coding. Also, whether P-SCO ratings or narrative comments changed resident behaviors was not measured.

Future research should study the P-SCO at multiple sites and examine to what extent the checklist leads to higher-quality comments that are aligned with the essential tasks of a medication visit, and, even more importantly, how the feedback can be used (eg, longitudinal coaching) to best support growth.

Conclusions

This study shows that a direct observation tool such as the P-SCO can yield high-quality narrative

feedback that addresses a variety of important competencies. Narrative comments not only add explanation and guidance for the high and low checklist scores, but also contribute information not conveyed by the checklist scores. The checklist provides the performance dimensions and frame of reference, while the comments provide the detail necessary for a learner to make changes and for a program to provide guidance.

References

1. Halman S, Dudek N, Wood T, Pugh D, Touchie C, McAleer S, et al. Direct observation of clinical skills feedback scale: development and validity evidence. *Teach Learn Med*. 2016;28(4):385–394. doi:10.1080/10401334.2016.1186552.
2. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;341:c5064. doi:10.1136/bmj.c5064.
3. Schuwirth LW, van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478–485. doi:10.3109/0142159X.2011.565828.
4. Al Ansari A, Ali SK, Donnon T. The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Acad Med*. 2013;88(3):413–420. doi:10.1097/ACM.0b013e318280a953.
5. Leep Hunderfund AN, Rubin DI, Laughlin RS, Sorenson EJ, Watson JC, Jones LK, et al. Validity and feasibility of the EMG direct observation tool (EMG-DOT). *Neurology*. 2016;86(17):1627–1634. doi:10.1212/WNL.0000000000002609.
6. Olupeliyawa AM, O'Sullivan AJ, Hughes C, Balasooriya CD. The Teamwork Mini-Clinical Evaluation Exercise (T-MEX): a workplace-based assessment focusing on collaborative competencies in health care. *Acad Med*. 2014;89(2):359–365. doi:10.1097/ACM.0000000000000115.
7. Watanabe Y, Bilgic E, Lebedeva E, McKendy KM, Feldman LS, Fried GM, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc*. 2016;30(3):832–844. doi:10.1007/s00464-015-4285-8.
8. Watson MJ, Wong DM, Kluger R, Chuan A, Herrick MD, Ng I, et al. Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia. *Anaesthesia*. 2014;69(6):604–612. doi:10.1111/anae.12625.
9. Feraco AM, Starmer AJ, Sectish TC, Spector ND, West DC, Landrigan CP. Reliability of verbal handoff assessment and handoff quality before and after implementation of a resident handoff bundle. *Acad Pediatr*. 2016;16(6):524–531. doi:10.1016/j.acap.2016.04.002.
10. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med*. 1995;123(10):795–799. doi:10.7326/0003-4819-123-10-199511150-00008.
11. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009;302(12):1316–1326. doi:10.1001/jama.2009.1365.
12. Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using In-Training Evaluation Report (ITER) qualitative comments to assess medical students and residents: a systematic review. *Acad Med*. 2017;92(6):868–879. doi:10.1097/ACM.0000000000001506.
13. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816–822. doi:10.1111/j.1365-2923.2008.03105.x.
14. Lockyer JM, Sargeant J, Richards SH, Campbell JL, Rivera LA. Multisource feedback and narrative comments: polarity, specificity, actionability, and CanMEDS roles. *J Contin Educ Health Prof*. 2018;38(1):32–40. doi:10.1097/CEH.0000000000000183.
15. Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617–1621. doi:10.1097/ACM.0000000000001669.
16. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296–306. doi:10.1111/medu.12637.
17. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539–1544. doi:10.1097/ACM.0b013e3182a36c3d.
18. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. 2017;51(4):401–410. doi:10.1111/medu.13158.
19. Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. *J Gen Intern Med*. 2004;19(5 pt 2):558–561. doi:10.1111/j.1525-1497.2004.30134.x.
20. Pelgrim EA, Kramer AW, Mokkink HG, Van der Vleuten CP. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. *BMC Med Educ*. 2012;12:97. doi:10.1186/1472-6920-12-97.

21. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2016;91(10):1359–1369. doi:10.1097/ACM.0000000000001175.
22. Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM. Mixed messages or miscommunication? Investigating the relationship between assessors' workplace-based assessment scores and written comments. *Acad Med*. 2017;92(12):1774–1779. doi:10.1097/ACM.0000000000001743.
23. Cheung WJ, Dudek NL, Wood TJ, Frank JR. Supervisor-trainee continuity and the quality of work-based assessments. *Med Educ*. 2017;51(12):1260–1268. doi:10.1111/medu.13415.
24. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. *Med Educ*. 2015;49(9):909–919. doi:10.1111/medu.12777.
25. Young JQ, Irby DM, Kusz M, O'Sullivan PS. Performance assessment of pharmacotherapy: results from a content validity survey of the Psychopharmacotherapy-Structured Clinical Observation (P-SCO) Tool. *Acad Psychiatry*. 2018;42(6):765–772. doi:10.1007/s40596-017-0876-0.
26. Young JQ, Lieu S, O'Sullivan P, Tong L. Development and initial testing of a structured clinical observation tool to assess pharmacotherapy competence. *Acad Psychiatry*. 2011;35(1):27–34. doi:10.1176/appi.ap.35.1.27.
27. Young JQ, Rasul R, O'Sullivan PS. Evidence for the validity of the Psychopharmacotherapy-Structured Clinical Observation Tool: results of a factor and time series analysis. *Acad Psychiatry*. 2018;42(6):759–764. doi:10.1007/s40596-018-0928-0.
28. Young JQ, Hasser C, Hung EK, Kusz M, O'Sullivan PS, Stewart C, et al. Developing end-of-training entrustable professional activities for psychiatry: results and methodological lessons. *Acad Med*. 2018;93(7):1048–1054. doi:10.1097/ACM.0000000000002058.
29. Archer JC. State of the science in health professional education: effective feedback. *Med Educ*. 2010;44(1):101–108. doi:10.1111/j.1365-2923.2009.03546.x.
30. Sabey CV, Charlton C, Charlton SR. The "magic" positive-to-negative interaction ratio: benefits, applications, cautions, and recommendations. *J Emotional Behav Dis*. 2019;27(3):154–164.
31. Young JQ, Nelson JC. Reconceptualizing medication management: implications for training and clinical practice. *J Clin Psychiatry*. 2009;70(12):1722–1723. doi:10.4088/JCP.09ac05828whi.
32. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: rater performance and behavior when considering multiple competencies. *Teach Learn Med*. 2016;28(1):41–51. doi:10.1080/10401334.2015.1107489.
33. Lefebvre C, Hiestand B, Glass C, Masneri D, Hosmer K, Hunt M, et al. Examining the effects of narrative commentary on evaluators' summative assessments of resident performance. [published online ahead of print December 26, 2018.] *Eval Health Prof*. doi:10.1177/0163278718820415.
34. 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.

John Q. Young, MD, MPP, PhD, is Vice Chair for Education, and Professor, Department of Psychiatry, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell; **Rebekah Sugarman, AB,** is Research Coordinator, Department of Psychiatry, Zucker Hillside Hospital, Northwell Health; **Eric Holmboe, MD,** is Senior Vice President, Accreditation Council for Graduate Medical Education; and **Patricia S. O'Sullivan, EdD,** is Professor, Department of Medicine, and Director of Research and Development in Medical Education, University of California, San Francisco School of Medicine.

Funding: This study was funded by the American Board of Psychiatry and Neurology, Research Award (2017–2018).

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

Portions of the data in this manuscript were previously presented at the 2nd World Summit on Competency Based Medical Education during the Association for Medical Education in Europe (AMEE) Annual Conference, Basel, Switzerland, August 24–29, 2018 and the AMEE Annual Conference, Vienna, Austria, August 25–28, 2019.

Corresponding author: John Q. Young, MD, MPP, PhD, Zucker School of Medicine, Department of Psychiatry, Kaufman 217a, 75–59 263rd Street, Glen Oaks, NY 11004, 718.470.8005, jyoung9@northwell.edu

Received March 23, 2019; revisions received July 7, 2019, and August 3, 2019; accepted August 5, 2019.