Assessors' interpretations of narrative data on communication skills in a summative OSCE

Assessors' interpretations of narrative data on communication skills in a summative OSCE

Kyle John Wilby,[1] Diana H J M Dolmans,[2] Zubin Austin[3] & Marjan J B Govaerts[2]

OBJECTIVES Increasingly, narrative assessment data are used to substantiate and enhance the robustness of assessor judgements. However, the interpretation of written assessment comments is inherently complex and relies on human (expert) judgements. The purpose of this study was to explore how expert assessors process and construe or bring meaning to narrative data when interpreting narrative assessment comments written by others in the setting of standardised performance assessment.

METHODS Narrative assessment comments on student communication skills and communication scores across six objective structured clinical examination stations were obtained for 24 final-year pharmacy students. Aggregated narrative data across all stations were sampled for nine students (three good, three average and three poor performers, based on communication scores). A total of 10 expert assessors reviewed the aggregated set of narrative comments for each student. Cognitive (information) processing was captured through think-aloud procedures and verbal protocol analysis.

RESULTS Expert assessors primarily made use of two strategies to interpret the narratives, namely comparing and contrasting, and forming mental images of student performance. Assessors appeared to use three different perspectives when interpreting narrative comments, including those of: (i) the student (placing him- or herself in the shoes of the student); (ii) the examiner (adopting the role of examiner and reinterpreting comments according to his or her own standards or beliefs), and (iii) the professional (acting as the profession's gatekeeper by considering the assessment to be a representation of real-life practice).

CONCLUSIONS The present findings add to current understandings of assessors' interpretations of narrative performance data by identifying the strategies and different perspectives used by expert assessors to frame and bring meaning to written comments. Assessors' perspectives affect assessors' interpretations of assessment comments and are likely to be influenced by their beliefs, interpretations of the assessment setting and personal performance theories. These results call for the use of multiple assessors to account for variations in assessor perspectives in the interpretation of narrative assessment data.

Medical Education 2019: 53: 1003–1012
doi: 10.1111/medu.13924

[1]School of Pharmacy, University of Otago, Dunedin, New Zealand
[2]School of Health Professions Education (SHE), Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands
[3]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Ontario, Canada

Correspondence: Kyle John Wilby, School of Pharmacy, University of Otago, PO Box 56, Dunedin 9054, New Zealand.
Tel: 00 64 3 479 7325; E-mail: kyle.wilby@otago.ac.nz

INTRODUCTION

There are increasing calls to capture detailed descriptive performance information for trainee assessment[1–3] and narrative assessment comments have been advocated for use within a variety of assessment contexts, including objective structured clinical examinations (OSCEs).[4–6] Narrative data provide the ability to detect student- and context-specific areas of strength or weakness in task performance, and may thus support judgement and decision making by capturing data that may be lost when using checklists, scales or rubrics.[5,7,8] As performance assessments require assessors to observe and interpret students' performance of professional tasks, narrative assessment data may furthermore make explicit assessors' reasoning in data interpretation and judgement.[3,4] Use of narrative assessment data is therefore quickly gaining credibility across assessment settings and, in the era of competence-based and programmatic assessment, assessors are increasingly tasked to interpret these qualitative assessment data separately from numeric performance scores.[1,8]

Although studies in workplace settings have shown that narrative assessment data can substantiate assessor judgements and may be reliable, as well as discriminatory of student performance, little is known about how assessors interpret and bring meaning to narrative comments in standardised assessment contexts.[4,9] Research on workplace-based assessments has shown that the interpretation of narrative data written by others may be challenging.[7,10] Research findings indicate that the language used by assessors in narratives is often vague and generic; assessors use 'hidden codes' and other linguistic strategies that must be deciphered in order to understand the intended meaning of comments.[7,11] Additionally, studies show that factors such as assessor expertise, assessor beliefs and assessor perceptions of the assessment tasks and assessment context may influence the types of data assessors consider relevant in performance assessment or the ways in which assessors interpret descriptive assessment data.[12–14] Differing beliefs about the assessment task, for example, may result in stricter or more lenient judgements depending on the assessor's perceptions of how closely the assessment relates to real-life professional practice.[15] It has been argued that variations in these judgements are likely to be less about what assessors focus on or pay attention to when reviewing data and more about how they bring meaning to the data through the development of a coherent representation or story.[10] It is the complexity of this process and the diverse nature of the data that make it difficult for assessors to reliably interpret comments on performance.

Clearly, given increasing calls for the use of narrative comments in performance assessments, we need to gain a better understanding of how assessors conceptualise and bring meaning to written assessment comments. Understanding assessors' processing in the interpretation of such data may aid the design of robust assessment approaches that incorporate narrative to support judgement and decision making. The purpose of this study was to explore how expert assessors process and bring meaning to narrative data when interpreting narrative assessment comments written by others in the setting of standardised performance assessment.

METHODS

Study design

This was a qualitative study conducted using a case study approach, in which we considered every individual assessor reviewing a set of narratives to represent a case. We used a think-aloud procedure[16] and verbal protocol analysis to capture how assessors brought meaning to narrative comments.

Setting

The study was conducted at the College of Pharmacy at Qatar University. The College maintains a Bachelor of Science in Pharmacy programme accredited by the Canadian Council for Accreditation of Pharmacy Programs (CCAPP). The programme graduates approximately 25 female students per academic year. All students complete a summative OSCE at the end of the 4-year curriculum. Previous studies have shown the psychometrics of this OSCE to be acceptable for summative assessment.[17]

Participants

A total of 10 expert assessors were recruited to review the aggregated narrative datasets obtained from the Year-4 summative OSCE. These expert assessors were pharmacists, had current or recent (within 3 years) practice experience, and had previously been trained for and evaluated student

communication skills during OSCEs. Expert assessors were purposively sampled from the assessor pool at Qatar University according to these criteria. No further training was provided to assessors in this study beyond the provision of instructions for the think-aloud procedure. No compensation for participation was provided.

Research procedures

Step 1: collection of narrative assessment data in a summative OSCE

For the present study, we used assessment data from the OSCE conducted in 2018. A total of 24 graduating pharmacy students completed the OSCE, which included six 8-minute communication-focused stations. Three stations required the student to interact with a standardised patient, two with a standardised physician, and one with a standardised mother. A blueprint of the stations is provided in Table 1. Six OSCE examiners (one per station) were recruited to write narrative evaluations of students' communication skills and to score communication skills according to a single-dimension 5-point rating scale anchored at 1, 3 and 5 points (1 = communicates inappropriately and ineffectively for the task; 3 = communicates with some logic and comprehension, but not consistently; 5 = communicates precisely, logically and perceptively for the encounter, integrating all relevant components).[17,18] All examiners had been previously trained using the tool via pre-assessment exercises and post-assessment debriefing. In addition, all examiners had undergone the same previous training and had experience in writing narratives in a past OSCE assessment. Upon completion of the OSCE, the six narrative evaluations obtained for each student were de-identified and compiled as a set for analysis.

Step 2: review of narrative data sets by expert assessors (think-aloud procedure)

A pilot procedure was conducted with two eligible assessors (results not included in the analysis). This aimed to determine the number of aggregated student narrative sets that could realistically be reviewed during study procedures without placing excessive burden on participants. It was determined that eight to 10 sets were optimal and allowed for the completion of the study protocol within 2 hours. Based on this result and previous research showing that narratives can discriminate between performance levels,[19,20] we therefore purposively sampled three sets of narrative evaluations from groups of students representing low (mean ± standard deviation [SD] score: 2.8 ± 0.99), average (mean ± SD score: 3.9 ± 0.96) and high (mean ± SD score: 4.7 ± 0.57)

Table 1 Blueprint of the objective structured clinical examination

Station	Standardised actor	Description
1	Patient	A young adult female presents to a community pharmacy for an oral contraceptive. The pharmacist must recognise a drug interaction and recommend an appropriate barrier method
2	Physician	A physician presents a patient to the pharmacist in a primary care setting. The pharmacist must recognise the need for and recommend a renal dose adjustment
3	Physician	A physician approaches a pharmacist in a hospital setting for prescribing advice for a pneumonia patient. The pharmacist must educate the physician regarding appropriate antibiotic step-down therapy
4	Patient	A patient with a language barrier presents to a community pharmacy with heartburn. The pharmacist must communicate, using non-verbal techniques (i.e. pictograms) instructions for non-prescription medications and self-care
5	Patient	A patient presents to a community pharmacy with acute shortness of breath and has risk factors for pulmonary embolism. The pharmacist must urgently refer the patient to emergency services
6	Paediatric patient's mother	A mother is picking up an antibiotic prescription for her child. The pharmacist must provide appropriate counselling and instructions

performers based on the overall scores provided by the OSCE examiners. This sampling strategy was selected simply to ensure that there would be (theoretically) enough variation in the narrative data for interpretation.

Each expert assessor was scheduled for a meeting with the same investigator with the purpose of completing a think-aloud protocol.[16] Assessors were briefed on the study objectives and procedures and asked to provide written and signed informed consent, including permission for audiotaping. The interviewer then provided the assessor with an aggregated set of narrative evaluations (without communication scores) corresponding to one student. The order of student datasets presented to each assessor remained constant for the study in order to simulate assessment approaches in real life. The assessor was asked to begin reading the narratives and to verbalise all of his or her thoughts as they emerged when reading and interpreting assessment data. The investigator prompted the assessor by saying 'Please continue thinking aloud' if the assessor ceased to verbalise his or her thoughts for more than a few seconds. When the assessor signalled that he or she had finished reviewing the aggregated data for the student, the investigator ceased prompting. The same procedure was repeated for each of the nine student narrative datasets. As the purpose of this study was to determine how assessors process and give meaning to narrative data, assessors were not required to make an overall judgement of performance or to provide an overall performance score. Field notes were taken to help inform data analysis because they capture actions that would not be analysable from the transcript alone (e.g. facial expressions, gestures).

Data analysis

All think-aloud protocols were transcribed verbatim by one investigator (KJW), who read all transcripts multiple times before coding. Transcripts were coded inductively by two independent coders (KJW and a research assistant). Codes were assigned using an open coding procedure that identified segments that related to the research question, and that reflected assessors' approaches to interpreting and bringing meaning to the narratives. Coding was thus based on *how* assessors brought meaning to comments, rather than on the specific judgements they made. For example, a remark such as 'I see a pattern in comments across stations that the student had good non-verbal communication and is a good communicator' would be coded as representing 'comparison across stations' rather than 'good communication'. Coders compared and discussed codes throughout the coding process to clarify discrepancies. Once all transcripts had been coded, codes were combined into broader themes, which were discussed amongst the research team. Subsequently, the same coders independently conducted a between-case analysis by comparing and contrasting patterns across cases (assessors) in order to search for similarities and differences in how assessors approached and interpreted narrative data. The results were discussed amongst the research team on multiple occasions until the final themes were agreed upon. Representative quotes were extracted from the transcripts to illustrate the themes.

RESULTS

We found that expert assessors used different strategies to bring meaning to aggregated narratives that helped them to 'paint a picture' or 'build a story' of student performance. Expert assessors compared and contrasted information within a student's narrative set and across narrative sets that pertained to different students. Assessors also engaged in creating mental images of what had happened during the OSCE in order to better understand or 'visualise' what had occurred during the interaction. Furthermore, assessors adopted different perspectives in the interpretation of narrative comments, resulting in different ways of explaining, reframing or understanding written comments. The data allowed us to construct three predominant perspectives used by assessors to bring meaning to comments: (i) the student perspective; (ii) the examiner perspective, and (iii) the professional perspective of real-life practice. In the following sections, we will present and discuss our findings relating to these strategies and perspectives in more detail. Examples of narrative comments obtained for each performance level are provided in Box 1.

Strategies used in the interpretation of narrative data

Comparing and contrasting

For each individual student, assessors compared and contrasted comments both within and across stations in order to seek patterns of student performance and to arrive at a coherent interpretation of the student's communication

Box 1 Examples of narrative comments

Example 1 (Student 1, Station 3):

- Great eye contact
- Good questions – clear
- Asks clarifying questions
- Positive feedback to good news (e.g. 'That's great the patient is improving')
- Leaning in, nodding/acknowledging responses from the physician
- Systematic process – asked most important questions first
- Clarified question/summarised patient characteristics
- Good respectful tone, clear voice, few interruptions (ums, ers)
- Provided rationale for recommendation, very clear

Example 2 (Student 4, Station 1):

- She was awkward in the introduction – waited for patient to say something
- Used too many 'uh's'
- Took watch off in middle of interaction
- Abrupt at times
- Said: 'How many times do you do intercourse?' (Not appropriate)
- Kept checking watch
- Hand placed by face (not confident)
- Provided good eye contact
- Asked assessor for time (broke interaction)
- Very rushed at end
- Did not communicate effectively and was potentially harmful to the patient

skills. The identification and confirmation of patterns of similarity across stations facilitated the development of a coherent, overall story of the student's performance. Inconsistencies or 'red flags' were also identified and acknowledged, yet were often incorporated into the overall story by looking for context-specific (or task-specific) explanations for the discrepancy:

> [In response to: *Confident when making recommendation*] Again Station 2, very consistent with Station 1. It seems the student was able to make a confident recommendation, suggesting that the student is listening, is able to process that information, and is able to give that information to a physician from a professional perspective, signifying that the physician is likely to trust that recommendation. (Assessor [A] 8, Student [S] 3)

> [In response to all comments for Student 1] Again she is making herself quite approachable and calm towards the patient so I don't think there would be any issues for this student except for Station 4, when she struggled with the language barrier. (A1, S1)

Once they had reviewed data for several students, assessors also began to compare between students to further confirm their developing story:

> [In response to: *Confident in tone and expressions; finished interaction confidently; is very professional*] So similar to Student 3, where I said it was a strong student, I am again getting that picture with this one. Someone sitting down and really taking command of the interaction. (A4, S6)

Formation of mental images

Assessors appeared to build their stories of student performance by forming mental images of student–patient interactions according to the descriptions provided within the narrative comments. Assessors verbalised their images but also used details of hand gestures and facial expressions during the think-aloud protocol to bring meaning to what had been written:

> [In response to: *Smiled with nice greeting, which helped to develop rapport; eye contact reasonable; body position showed command of the interaction; lots of laughing and smiling – may influence the patient's perception of professionalism*] How I'm making sense of this is we have a patient that is coming in for a standard drug-related problem, the student has developed an initial rapport, smiled, greeted. So if I was to kind of visualise how this student is acting, she has good body position, is making lots of eye contact, is laughing and smiling and I would think it is more of making a patient at ease. (A8, S1)

The detailed nature of the data and the presence of the examiner's interpretation of performance facilitated the assessor's ability to form mental images of the student's communication behaviours. When assessors did not have enough detail within the narrative to understand what had occurred during the interaction (i.e. when assessors were not able to visualise), they had difficulty in interpreting

narratives and at times appeared to ignore or disregard these comments:

> [In response to: *Excellent verbal; respectful; systematic*] Hard to do much with this. It seems all positive but I can't visualise this and I'm not getting a sense of what happened throughout the interaction. I'm just getting words of 'respectful'. OK, in what context? What was 'respectful' about it? I have no idea. (A4, S9)

Assessors' perspectives in the interpretation of narrative data

Data analysis showed that assessors focused on and provided interpretations for many of the same comments during the think-aloud procedure. The analysis of think-aloud protocols resulted in the identification of three different perspectives that assessors seemed to adopt when reviewing written comments: the student perspective; the examiner perspective, and the professional perspective of real-life practice. Table 2 provides illustrative examples of specific comments that show how the perspective brought to the comments by the assessor greatly influenced the meaning interpreted.

Student perspective

Assessors who took the student perspective when interpreting comments attempted to understand why a student might have exhibited the behaviours documented within the narrative comments. These assessors empathised with the student and tended to relate negative behaviours to factors beyond the student's control. Assessors appeared to search for reasons or justifications for student behaviours, rather than to accept them according to the face value of how the comments were written:

> [In response to: *Pauses during questioning, appears a bit unsure*] I think when a student pauses during the interaction they are thinking about what to say or trying to just recap the information in their mind. (A3, S2)

Table 2 Illustrative examples of assessors' perspectives of the same narrative comments

Narrative comment	Response from the 'student perspective'	Response from the 'professional perspective'	Response from the 'examiner perspective'
Didn't use the pictograms for all instructions although they were available	Maybe the students are so focused they don't see what is available for them (A3)	But definitely the negative is she didn't use the pictograms even though they were available. So overall her communication skills were not as good in that case (A9)	That could be interpreted as being a positive as they may have used other ways and didn't just go to the pictograms. I'm seeing the assessor sees it as wrong but I'm fine as long as the patient seems to gather an understanding (A4)
Major heavy breathing when checking references	Ahh that is interesting. I've never came across a student who would be talking and had major heavy breathing. Is she frustrated with this station? Maybe she is feeling uncomfortable that she has to talk about this topic (A3)	Whenever you have heavy breathing when checking references it means that you are nervous. If you are nervous and you lack self-confidence, how can you get the patient to trust you or be interested in communicating with you? (A10)	I can see that the pharmacist is anxious (A5)
Took watch off in middle of interaction	I think some people when they get nervous they take off their accessories for some reason. I think it was fine if the student was able to maintain their posture (A2)	So took the watch off? Anxious about the examination and not very focused can lose the attention from or rapport with the patient (A9)	I will assume it is removing someone's watch, maybe that is what the examiner means (A6)

A = assessor.

> [In response to: *Eye contact reasonable – looked at notes throughout though*] This may be caused by the nervousness. The student was a little bit nervous so she was trying to hold notes or something to relieve the stress or nervousness. (A7, S1)

Examiner perspective

Assessors who assumed the perspective of the examiner attempted to understand the meaning of comments by placing themselves in the shoes of the examiner who had been present. They appeared to accomplish this by assuming the role of examiner when visualising what had occurred within the interaction. Commonly, they specifically mentioned or alluded to the examiner, rather than the student, in their comments and interpretations. In general, these assessors showed scepticism towards the written comments and were quick to provide different opinions with respect to how the OSCE examiner appeared to have interpreted student performance:

> [In response to: *Said 'how many times do you do intercourse?' (not appropriate)*] [The examiner] said [this was] not appropriate. I don't see this as not appropriate. (A5, S4)

> [In response to: *Did not use the pictograms all the time (sign language, facial expressions, etc . . .) . . . good but was easier for the patient to understand pictograms*] Again, that could be interpreted as being a positive . . . I'm seeing the [examiner] sees it as wrong . . . when I read the first part and they are using sign language and facial expressions, I'm fine with that. (A4, S6)

> [In response to: *In general tone was not warm but polite (more monotone)*] So what is the meaning [of] the tone was not warm? Does [the student] need to hug the doctor or what? (A5, S14)

Professional perspective of real-life practice

Assessors who assumed the professional perspective interpreted comments by relating meaning to what would occur in 'real life' and how the student's behaviour might have impacted the patient. From this perspective, assessors commonly identified the student as a 'pharmacist'. They identified important considerations relating to culture and trust, both in the context of encompassing the role of a good professional. They did not tolerate unprofessional behaviour and commonly reflected on how negative behaviours (i.e. red flags) may be detrimental to patient care or professional relationships:

> [In response to all comments for Station 4] It is like she had no idea what to do. You can't act like that in a professional setting as typically you are the only pharmacist available. So, if you did that it is not good, it could be detrimental to a patient's health. (A1, S4)

> [In response to all comments for Station 1] This, from a patient's perspective, really limits the professionalism of the student and [the patient] would probably not end up trusting a recommendation in the end. (A8, S2)

> [In response to: *Not linking information provided by the patient very well, which resulted in repeating questions sometimes*] I'm quite sure it is not the right time to repeat questions and waste time because it is an emergency. The pharmacist should be focused to save time and not waste time. (A10, S16)

DISCUSSION

The purpose of this study was to explore how expert assessors process and bring meaning to narrative data when interpreting narrative assessment comments written by others in the setting of standardised performance assessment. We found that assessors gave meaning by comparing and contrasting written comments within and across sets of student data to search for performance patterns, and by using the narrative to construct a mental image of what had occurred during the interaction. In addition, assessors seemed to take different perspectives when interpreting assessment comments (i.e. the perspective of the student [placing him- or herself in the shoes of the student], the perspective of the examiner [reinterpreting comments according to his or her own standards or beliefs], or the perspective of the professional [acting as the profession's gatekeeper by protecting the patient and considering the assessment as a representation of real-life practice]). Our findings show that these differences in assessor perspectives may affect assessors' understandings and explanations of assessment comments. As such, our findings may contribute to further understanding of assessor variability when such data are used to support judgements and in decision making.

In line with previous research, our findings suggest that assessors' approaches to and the perspectives from which they engage in the interpretation of narrative data are influenced by their beliefs, interpretations of the assessment setting and personal performance theories.[12–14] For example, an assessor who assumes the perspective of the student may do so because he or she believes that the assessment task and the setting itself may affect a student's performance either positively or negatively. Assessors who take the examiner perspective (and doubt others' comments) may do so because they hold different performance standards and conceptualisations of what is to be considered effective task performance behaviour. Assessors who assume the perspective of the professional may do so because they feel their primary role is to protect the public and to act as gatekeepers to the profession and because they believe the OSCE is an authentic representation of real-life situations and that it allows for the extrapolation of assessment results to professional practice. Based on these findings, it is conceivable that different assessor perspectives may result in variability in performance judgement and decision making.[15] For instance, the assessor who takes the student perspective may be more lenient (forgiving) in comparison with the assessor who takes the professional perspective and considers him- or herself as the profession's gatekeeper and hence feels a need to be very strict. However, further research is required to investigate the extent to which assessors' perspectives actually influence performance judgements when narratives derived from the context of a summative OSCE or other standardised assessment are used.

Our findings are also in line with previous research showing that assessor variance may not result from what assessors focus on, but, instead, may be attributable to the ways in which assessors interpret and bring meaning to data.[10,12,13] Our study adds to this argument by identifying and describing two different strategies and three different perspectives that assessors may adopt when conceptualising student performance based on narrative assessment data sourced from a high-stakes assessment context. It appears that the complexity of how assessors approach and process these data (e.g. from different perspectives) may be the potential driver for variability in judgement. Although our data suggest that every assessor in our study largely interpreted narrative assessment comments through the adoption of a predominant perspective, further research is needed to investigate if and when, or under which conditions, assessors switch perspectives.

Limitations

This study should be interpreted in the light of some limitations. Firstly, this was a single-centre study and hence the types of comment, context of assessment and experience of assessors may limit the transferability of results. It should be noted, however, that all assessors were experienced and purposively selected based on their previous participation in OSCE assessment. Secondly, examiners were instructed to write comments as if they would be used for assessment decisions, yet were aware that the comments would not be used for grading purposes. This may have affected the language or amount of detail examiners chose to provide. Thirdly, we presented the narrative comments to each assessor in the same way each time, which may have introduced bias from order effects. Although this may have elicited some different impressions or reactions on the part of the assessors when comparing and contrasting or visualising the data, we did it deliberately to reflect how assessors would be likely to receive a set of narrative comments in real-life practice and we believe that any effect would impact our findings only negligibly. Despite these limitations, we feel our findings are important for elucidating how assessors interpret and use narrative data, and help to better clarify assessor variability in performance assessments.

Implications

This study builds on previous research relating to the interpretation of descriptive performance data and has implications for assessment practice. Our findings support arguments that the key to credible assessment does not lie in stripping or reducing data into smaller or simplified components, or in standardisation or assessor training, and that we must develop ways to cope with the variance and ambiguity that may arise from differing interpretations.[10] Previous suggestions for the formation of clinical competency committees, on which multiple assessors review data to make performance decisions, may indicate a good way of starting to account for the variation we (and others) have observed. The involvement of multiple assessors at all stages in the interpretation and judgement of performance is necessary to ensure robustness of judgement, but may also increase the

richness of assessment data for the purposes of understanding student performance across assessment settings. Including data obtained from standardised patients may also enhance the quality of performance assessments, although further research is needed to investigate if and how these data contribute to assessors' interpretations of performance.

Our results support the need to pay attention to the factors related to assessor variance in assessor training. Despite the limited nature of what formal training can achieve, recognising assessor beliefs and performance theories and accounting for these in coaching or providing feedback may promote awareness of the strategies and perspectives used by assessors in judging performance. Furthermore, our findings show that assessors continue to compare and contrast students relative to one another. This implies that not only should OSCE examiners be trained and coached to write meaningful narratives that enable the evaluation of student performance against performance standards, but that assessors should also be trained in the use of performance data to ensure performance assessment is actually criterion based (rather than normative based). Finally, based on the results of this study, the exploration of how assessors' interpretations influence overall conclusions (and any relationship with scores) is an obvious next step for future research.

CONCLUSIONS

This study explored how assessors process narrative data obtained from summative OSCEs. Assessors bring different perspectives to narrative comments, which appear to influence their interpretations of assessment data. These findings support the notion that assessor variance may be the result of many factors working collectively during the assessment task, including how assessors approach and interpret descriptive data. The results from our study can be used to enhance our understanding of the assessment process in order to inform the development and refinement of assessment procedures for the collection and interpretation of narrative performance data. Based on our results, we can conclude that multiple assessors should interpret narrative data in order to account for variation in assessors' approaches and perspectives.

Contributors: KJW contributed to the conception and design of the study, conducted the procedures, analysed the data, interpreted the results and drafted the first version of the manuscript. DHJMD, ZA and MJBG contributed to the conception and design of the study, and interpreted the results obtained. All authors (KJW, DHJMD, ZA an MJBG) contributed to the critical revision of the manuscript, approved the final version for publication and have agreed to be accountable for all aspects of the work.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: this study was approved by the Qatar University Institutional Review Board (QU-IRB 942-EA/18).

REFERENCES

1 Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39**:309–17.
2 Eva KW. Cognitive influence on complex performance assessment: lessons from the interplay between medicine and psychology. *J Appl Res Mem Cogn* 2018;**7** (2):177–88.
3 Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol* 2013;**4**:668.
4 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):1617–21.
5 Van Nuland M, van den Noortgate W, van der Vleuten C, Goedhuys J. Optimizing the utility of communication OSCEs: omit station-specific checklists and provide students with narrative feedback. *Patient Educ Couns* 2012;**88** (1):106–12.
6 Harrison CJ, Molyneux AJ, Blackwell S, Wass VJ. How we give personalised audio feedback after summative OSCEs. *Med Teach* 2015;**37** (4):323–6.
7 Ginsburg S, Regehr G, Lingard L, Eva K. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ* 2015;**49** (3):296–306.
8 Driessen E, van der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ* 2005;**39** (2):214–20.
9 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):1539–44.
10 Schutz A, Moss PA. Reasonable decisions in portfolio assessment: evaluating complex evidence of teaching. *Educ Policy Anal Arch* 2004;**12**:33.
11 Ginsburg S, van der Vleuten C, Eva KW, Lingard L. Hedging to save face: a linguistic analysis of written

comments on in-training evaluation reports. *Adv Health Sci Educ Theory Pract* 2016;**21** (1):175–88.
12 Govaerts MJB, van de Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract* 2013;**18** (3):375–96.
13 Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. *Adv Health Sci Educ Theory Pract* 2018;**23** (2):275–87.
14 Berendonk C, Stalmeijer R, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ Theory Pract* 2013;**18** (4):559–71.
15 McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk–dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modeling. *BMC Med Educ* 2006;**6**:42.
16 Van Semeren MW, Barnard YF, Sandberg JAC. *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. London: Academic Press 1994.
17 Sobh AH, Austin Z, Izham MI, Diab MI, Wilby KJ. Application of a systematic approach to evaluating psychometric properties of a cumulative exit-from-degree objective structured clinical examination (OSCE). *Curr Pharm Teach Learn* 2017;**9** (6):1091–8.
18 Munoz LQ, O'Byrne C, Pugsley J, Austin Z. Reliability, validity, and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy. *Pharm Educ* 2005;**5** (1):1–12.
19 Wilby KJ, Govaerts M, Austin Z, Dolmans D. Discriminating features of narrative evaluations of communication skills during an OSCE. *Teach Learn Med* 2019;**31** (3):298–298.
20 Wilby KJ, Govaerts MJB, Dolmans DHJM, Austin Z, van der Vleuten C. Reliability of narrative assessment data on communication skills in a summative OSCE. *Patient Educ Couns* 2019;**102** (6):1164–69.

Received 30 January 2019; editorial comments to authors 8 March 2019; accepted for publication 29 May 2019