When numbers become words

When numbers become words:
Assessors' processing of performance data within OSCEs

The research reported here was carried out at

In the School of Health Professions Education

Cover illustration by Kyle John Wilby
Printing:
ISBN:

When numbers become words:
Assessors' processing of performance data within OSCEs

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht, op gezag van de Rector Magnificus, Prof. dr. Rianne M. Letschert, volgens het besluit van het College van Decanen, in het openbaar te verdedigen op

dinsdag 2 juli 2019, om 10.00 uur

door

Kyle John Wilby

Supervisor
Prof dr. D.H.J.M. Dolmans

Co-Supervisors
Dr. M.J.B. Govaerts
Prof dr. Z. Austin, University of Toronto, Toronto, Canada

Assessment Committee
Prof dr. J. Rethans (Chairman)
Prof dr. P. Brand, Universitair Medisch Centrum Groningen, Groningen, Netherlands
Prof dr. S. Heeneman
Prof dr. M. van Nuland, Katholieke Universiteit, Leuven, Belgium
Dr. M. Verheggen

Table of Contents

Chapter 1

Introduction

Chapter 1

Introduction

The rise of competency-based education in health professionals' training programs, which aims to produce graduates who have the ability to integrate communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in day-to-day professional practice has changed the way we assess our students.[1,2] Traditionally used assessment approaches, such as multiple-choice tests, essays, and oral examinations, commonly focus on testing of factual recall and/or application of knowledge (cognitive tests). However, these methods do not allow for assessment of professional competence, i.e. the use of knowledge, skills, and attitudes in performance of professional tasks.[2,3] Performance assessments, on the other hand, allow educators to use authentic tasks to monitor and assess progress towards professional competence and to assess how students accomplish complex skills or tasks in professional settings.[3] Performance assessments, by assessing what students actually do versus assessing what they say they will do, are essential in making decisions about whether students are fit for practice, and specifying / identifying any gaps between actual performance and what is expected in high quality care.[4] These assessments have therefore become well established within health professionals' training programs.

A second change in current assessment practices is reflected in a shift towards capturing rich performance data, rather than solely relying on numbers and scores to make assessment decisions.[1] Rich data in the form of narrative assessment comments, allow for provision of meaningful feedback and information for learning, which might support student development and growth.[5,6] The use of narrative assessment data may furthermore ensure robustness of high-stakes decision making by substantiating judgements on task performance. As performance assessments require assessors (human judges) to observe and interpret students' performance of professional tasks, narrative assessment data may make assessors' reasoning in judgement and decision explicit.[5,7] Use of narrative assessment data is therefore quickly gaining credibility across assessment settings.[1,8]

Objective structured clinical examinations (OSCEs) are performance-based assessment formats in health professionals' education.[2,9] OSCEs are simulation-based interactions that typically utilize standardized patients to present the same case in the same way (i.e. standardized) for all students. OSCEs are used to assess a sample of the skills that learners are expected to have mastered, yet most often include stations to assess clinical skills such as history-taking and patient assessment skills as well as communication with patients.[10] Formatively, OSCEs can be used for development of competencies through feedback and reflection.[11] It is therefore obvious that narrative data may have a role in OSCE assessment. From a summative perspective, OSCEs are used to 'objectively' assess a student's competency or competencies through demonstration of completion of a series of standardized tasks.[12] Checklists, rating scales, or rubrics are commonly used to guide assessors in the assessment, requiring assessors to convert their observations and judgments into scores and/or quantitative assessment data. In addition, assessors may be required to provide an overall global rating.[13] The purported objectivity of OSCEs is believed to result in a robust and defensible assessment.[9,12] As such, OSCEs are widely used as high stakes performance-based assessments to help determine if a student is ready to progress within a program, enter clinical training, or is safe for practice.[14]

Despite great effort to optimize the reliability of OSCEs through standardization of assessment tasks and assessor training, research consistently shows that OSCEs may not be as reliable as we may think. Research findings show that reliability is affected by problems related to task- or content-specificity, necessitating inclusion of large numbers of stations.[15,16] Recent studies, however, also suggest that between-assessor differences may affect reliability of these types of performance assessments – despite assessors being well-trained.[1] Well-known and well-researched examples of assessor bias include leniency or strictness bias, contrast effects, recency bias or differences in conceptualizations of competent performance.[17-20] In fact, findings from assessor cognition research suggest that multiple factors, including assessors' own experiences and beliefs, may influence assessors' processing of performance data, depending on what assessors consider of importance to pay attention to and subsequently encode, interpret and value when making judgements about observed performance.[21,22] Studies have consistently shown that this information processing is inherently idiosyncratic, meaning assessors tend to conceptualize competencies in different ways and also vary in how they interpret, synthesize, and eventually judge performance.[22-24] The quantitative data generated from OSCEs using checklists and rubrics do not provide insight into assessor judgements nor do they help to substantiate assessors' use of performance data in judgement and decision-making.[5] In order to capture and better understand how assessors process performance information and to ensure more credible and defensible (robust) decision-making using OSCE data, we need to make assessors' reasoning explicit, implying a shift from an exclusive focus on numbers and scores towards collection of rich, narrative assessment data.

Although studies have shown that narrative assessment data, as explained above, can provide feedback for learning, substantiate assessor judgements and result in reliable performance-based decision-making, so far little is known about how the use of narrative data impacts assessment quality. Recent studies from workplace-based assessment have shown that narrative data may be both reliable and discriminatory of student performance.[7,25] Previous research in the same settings, however, also showed that interpretation of narrative performance data written by others may be challenging.[20,26,27] Research findings indicate that assessors' language 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.[27,28] These findings may have consequences for assessment reliability, if comments are to be used for judgement and decision-making.[7] Obviously, the use of narratives in other settings, such as OSCEs, calls for studies investigating if and how narrative assessment impacts interpretation of assessment data and utility of data for pass-fail decision-making. Given research findings in workplace settings, questions can be raised, for example, about the language that OSCE-assessors use to convey their judgements to others, how others interpret assessors' messages, or how use of narratives in these standardized assessment settings impacts on assessment reliability. Overall, a better understanding of similarities and differences in what OSCE-assessors pay attention to, how they synthesize observations into meaningful assessment data, and how rich performance information is interpreted is essential in informing how to best incorporate narrative into OSCE assessments.

Research purpose and research questions

This PhD thesis aims to further our understanding of the usefulness of narrative performance data in OSCEs. Our research questions were:

1. How do assessors process performance data when judging student communication performance in OSCEs?
    a. What do assessors pay attention to when judging student communication in OSCE stations and how is this influenced by assessor characteristics?
    b. How do assessors convey their observations and interpretations into narrative assessment data?

2. How does use of narrative assessment data impact assessment quality?
    a. How do assessors interpret narrative assessment data (provided by others)?
    b. How reliable are scores based on narrative data obtained from an OSCE, as compared to scores based on direct observation?

Research setting and context

All studies included in this dissertation were conducted at the College of Pharmacy at Qatar University and included data from assessors trained to assess communication skills within OSCEs. Three of the four studies included in this dissertation were based on iterations of a summative OSCE required of students to complete at the end of the Bachelor of Science in Pharmacy program. It is an exit-from-degree OSCE that is linked to a final capstone clinical course. Previous to the OSCE, students completed 24 weeks of structured experiential training in workplace settings. The OSCE consists of 8-10 interactive stations that are blueprinted based on the program's competency framework. All stations require students to demonstrate effective communication skills by interacting with a standardized patient or other healthcare professional to solve a case related to clinical pharmacy practice. Communication was assessed according to a global 5-point rating scale and/or the use of narrative comments, as described in each study. Further information about the content and procedures in design and analysis of the OSCE is provided within each chapter.

Qatar University hosts the only pharmacist-training program in the country and graduates, on average, 25-30 students from the Bachelor of Science in Pharmacy program annually. All students are female and originate from Qatar or other countries within the Middle East and North Africa. The Canadian Council for Accreditation of Pharmacy Programs (CCAPP) accredits both the Bachelor of Science in Pharmacy degree and a post-graduate Doctor of Pharmacy degree. As such, competency frameworks and teaching and assessment strategies are largely adapted from the Canadian context.

Thesis Overview

This thesis presents four empirical studies that aim to further our understanding of how assessors process performance data within OSCEs and when implementing narrative assessment processes for student communication skills. In Chapter Two, we first investigate through the lens of cultural dimensions theory what assessors pay attention to when observing standardized interactions and how their cultural orientations may influence their scores and focus. In Chapter Three, we investigate how assessors synthesize performance observations and formulate narrative comments (i.e. behavioural constructs, linguistic patterns) to distinguish between good and poor performers. In Chapter Four, we investigate how assessors interpret aggregated narrative data and strategies they use to bring meaning to comments obtained from an OSCE. Chapter Five takes a closer look at the quality of narratives by using generalizability theory to estimate the reliability of judgements (i.e. communication scores) made based on narrative, as opposed to those obtained from direct observation during the OSCE. In Chapter Six, we attempt to answer our research questions, discuss the theoretical and practical implications of our work, and note the strengths and limitations arising from our analyses. In Chapter Seven, the valorisation implications of this research are explored. Summaries in both English and Dutch follow.

Table 1. PhD thesis studies overview

Chapter	2	3	4	5
Research questions	How are cultural orientations associated with communication scores, as well as assessors' interpretation and evaluation of affective and instrumental communication behaviours, when assessing simulated interactions?	How do assessors distinguish between good and poor performers when writing narrative assessment comments of communication skills during an OSCE?	How do expert assessors interpret and bring meaning to aggregated narrative assessment data of student communication skills during a summative OSCE?	How reliable is narrative assessment data regarding student communication skills obtained from a summative OSCE and how does this reliability compare to reliability of communication scores obtained from direct observations?
Methodology/ design	Video-based stimulated recall with verbal protocol analysis	Qualitative study using constructivist grounded theory and content analysis using politeness theory	Qualitative study using constructivist grounded theory	Generalizability theory
Context and participants	25 expert OSCE assessors	14 graduating pharmacy students completing a summative OSCE 18 expert assessors recruited to write narrative comments	24 graduating pharmacy students completing a summative OSCE 10 expert assessors recruited to interpret aggregated narrative comments	14 graduating pharmacy students completing a summative OSCE 12 expert assessors recruited to write narrative comments 2 expert assessors recruited to score narrative comments
Data Sources	Stimulated recall interview transcripts based on three videotaped interactions of varying student performance	Narrative assessment comments from a summative OSCE written about communication skills by 18 expert assessors (2 per	Think-aloud protocol transcripts based on assessors' interpretation of aggregated narrative assessment	Direct observation communication scores and scores obtained from scoring narrative assessment

		station)	comments across 6 stations for 9 students completing a summative OSCE	comments from 6 communication-focused stations

References

1. Eva KW. Cognitive influence on complex performance assessment: Lessons from the interplay between medicine and psychology. J App Res Mem Cogn. 2018;7:177-88.

2. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356:387-96.

3. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357:945-9.

4. Eva KW, Bordage G, Campbell C, Galbraith R, Ginsburg S, Holmboe E, Regehr G. Towards a program of assessment for health professionals: training into practice. Adv Health Sci Educ 2016;21:897-913.

5. 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.

6. Van Nuland M, Van den Noorgate W, van der Vleuten C, Goedhuys J. Optimizing the utility of communication OSCEs: Omit station-specific checklists and provide students with narrative feedback. Pat Educ Couns 2012;88:106-12.

7. 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;92:1617-21.

8. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ 2016;50:511-22.

9. Plakiotis C. Objective structured clinical examination (OSCE) in psychiatry education: A review of its role in competency-based assessment. Adv Exp Med Biol 2017;988:159-80.

10. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. Med Teach 2018;40:1208-13.

11. Pugh D, Desjardins I, Eva K. How do formative objective structured clinical examinations drive learning? Analysis of residents' perceptions. Med Teach 2018;4045-52.

12. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Med J 2011;80:127-32.

13. Wilkinson T. How not to put the O into an OSCE. Perspect Med Educ 2018;7:28-9.

14. 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;5:1-12.

15. Marwaha S. Objective structured clinical examinations (OSCEs), psychiatry and the Clinical assessment of Skills and Competencies (CASC) Same evidence, different judgement. BMC Psychiatry 2011;11:85.

16. Roberts C, Newble D, Jolly B, Reed M, Hampton K. Assuring the quality of high-stakes undergraduate assessments of clinical competence. Med Teach 2006;28:535-43.

17. Yeates P, Moreau M, Eva K. Are examiners' judgements in OSCE-style assessments influenced by contrast effects? Acad Med 2015;90:975-80.

18. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. Med Teach 2015;37:81-5.

19. Stroud L, Herold J, Tomlinson G, Cavalcanti RB. Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. Acad Med 2011;86:S8-S11.

20. Yeates P, Sebok-Syer SS. Hawks, doves and rash decisions: understanding the influence of different cycles of an OSCE on students' scores using many facet rash modeling. Med Teach 2017;39:92-9.

21. Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. Adv in Health Sci Educ. 2018;23:275-87.

22. Berendonk C, Stalmeijer R, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. Adv in Health Sci Educ 2013;18:559-71.

23. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the black box differently: assessor cognition from three research perspectives. Med Educ. 2014;48:1055-68.

24. Govaerts MJB, van de Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ. 2013;18:375-96.

25. 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;88:1539-44.

26. Schutz A, Moss PA. Reasonable decisions in portfolio assessment: evaluating complex evidence of teaching. Educ Policy Analysis Archives 2004;12:33.

27. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ 2015;49:296-306.

28. 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 in Health Sci Educ. 2016;21:175-88.

Chapter 2

Exploring the influence of cultural orientations on assessment of communication behaviours during patient-practitioner interactions

Published as: Wilby KJ, Govaerts MJB, Austin Z, Dolmans DHJM. Exploring the influence of cultural orientations on assessment of communication behaviours during patient-practitioner interactions. BMC Med Educ 2017;17:61.

Abstract

Background
Research has shown that patients' and practitioners' cultural orientations affect communication behaviors and interpretations in cross-cultural patient-practitioner interactions. Little is known about the effect of cultural orientations on assessment of communication behaviors in cross-cultural educational settings. The purpose of this study is to explore cultural orientation as a potential source of assessor idiosyncrasy or between-assessor variability in assessment of communication skills. More specifically, we explored if and how (expert) assessors' valuing of communication behaviours aligned with their cultural orientations (power-distance, masculinity-femininity, uncertainty avoidance, and individualism-collectivism).

Methods
Twenty-five pharmacist-assessors watched 3 videotaped scenarios (patient-pharmacist interactions) and ranked each on a 5-point global rating scale. Videotaped scenarios demonstrated combinations of well-portrayed and borderline examples of instrumental and affective communication behaviours. We used stimulated recall and verbal protocol analysis to investigate assessors' interpretations and evaluations of communication behaviours. Uttered assessments of communication behaviours were coded as instrumental (task-oriented) or affective (socioemotional) and either positive or negative. Cultural orientations were measured using the Individual Cultural Values Scale. Correlations between cultural orientations and global scores, and frequencies of positive, negative, and total utterances of instrumental and affective behaviours were determined.

Results
Correlations were found to be scenario specific. In videos with poor or good performance, no differences were found across cultural orientations. When borderline performance was demonstrated, high power-distance and masculinity were significantly associated with higher global ratings ($r=.445$, and .537, respectively, $p<0.05$) as well as with fewer negative utterances regarding instrumental (task focused) behaviours ($r= -.533$ and $-.529$, respectively). Higher masculinity scores were furthermore associated with positive utterances of affective (socioemotional) behaviours ($r=.441$).

Conclusions
Our findings thus confirm cultural orientation as a source of assessor idiosyncrasy and meaningful variations in interpretation of communication behaviours. Interestingly, expert assessors generally agreed on scenarios of good or poor performances but borderline performance was influenced by cultural orientation. Contrary to current practices of assessor and assessment instrument standardization, findings support the use of multiple assessors for patient-practitioner interactions and development of qualitative assessment tools to capture these varying, yet valid, interpretations of performance.

Background

Patient-practitioner interactions are essential for the delivery and uptake of healthcare services worldwide.[1] Provision of diagnostic testing or treatments without effective and safe communication with patients leads to poor health outcomes and mistrust in health systems.[1,2] Many definitions of culture exist, however it can be broadly defined as groups of people sharing similar values and beliefs systems.[3,4] This may include similarities based on origin, gender, religion, sexuality, and socioeconomic status, among others. Cultural diversity, i.e. differing values and beliefs systems between individuals, is implicated as a contributing factor to communication and health disparities.[5] Given increasing globalization of health care (migration of health professionals and patients alike) and patient-practitioner communication taking place in multicultural settings, health professions education programs must teach and assess effective communication strategies that prepare students to work with patients from diverse ethno-cultural and linguistic backgrounds.[6]

There are numerous challenges associated with addressing the effects of globalization on teaching and assessment. First, globalization of healthcare and health education now requires teachers and learners to work within multicultural environments and perhaps more importantly, assessment has to take place in multicultural settings.[7] More specifically, assessment settings may consist of assessors from different cultural backgrounds practicing in countries and cultures where they have neither learned nor practiced before. Secondly, cultural adaptation of assessment instruments and frameworks may be required to fit the local contexts and/or assessors may be required to use instruments and frameworks that do not match their personal conceptualizations of what constitutes effective communication.[8] These considerations can greatly affect assessment processes and pose risks to validity without proper understanding of how these cultural factors must be accounted for.

A widely used theoretical model in cross-cultural research is Hofstede's cultural dimensions theory, which summarizes five domains, or 'dimensions', that attempt to account for a spectrum of values and beliefs relating to a particular culture.[9] The five cultural dimensions are power-distance, individualism-collectivism, masculinity-femininity, uncertainty avoidance, and Confucian dynamism or long-term orientation.[10] The definition and explanation of each dimension is summarized in Table 1.

Table 1. Cultural dimension definitions and characteristics[9]

Dimension	Explanation
Power-distance	Extent to which less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally (refers to family, school, community, workplace, etc.)
Uncertainty avoidance	Extent to which the members of a culture feel threatened by uncertain or unknown situations and avoidance of such situations (expressed through nervousness and a need for predictability/rules)
Masculinity-femininity	Extent to which dominant values in society are masculine (in masculine societies gender roles are distinct whereas is feminine societies gender roles overlap)
Individualism-collectivism	Extent of ties between individuals (ties are loose in

	individualistic societies but integrated and cohesive in collectivist societies)
Long-term orientation	Extent of long or short-term orientation in life (long-term oriented people value persistence, thrift, sense of shame, and order; short-term oriented people value personal steadiness and stability, protecting oneself, respect for tradition, and reciprocation of greetings, favours, and gifts)

Findings from a recent study by Meeuwesen and colleagues (2009) suggest that Hofstede's cultural dimensions explain communication preferences during patient-practitioner interactions across physicians and patients in Europe.[11] More specifically, patients' and physicians' cultural orientations influenced both instrumental (defined as orientation, psychosocial talk, asking questions, counselling) and affective (social talk, agreement, backchannelling) behaviours in patient-practitioner interactions. Instrumental behaviours can be seen as more task-focused, while affective behaviours relate more to socioemotional exchange.[12] Meeuwesen and colleagues found, for example, that practitioners in highly individualistic countries showed more affective behaviours and focused less on instrumental behaviours such as asking questions and counselling. Practitioners from more masculine vs. feminine countries equally showed more affective behaviours (social talk and agreements). Practitioners high on uncertainty avoidance, on the other hand, paid more attention to instrumental behaviours related to psychosocial talk, whereas those higher on power-distance gave more orientation to their patients (instrumental), gave more social talk (affective), yet less backchannelling (affective).[11] Although these findings show a relationship between cultural orientations and communication behaviours in practitioners, any link between assessment of these behaviours and cultural orientations is unknown. Based on the communication preferences associated with cultural orientations, one might expect assessors to favour communication behaviours that match those specific to their own cultural orientation.[13]

Drawing from Hofstede's cultural dimensions theory, the purpose of this study is to explore cultural orientation as a potential source of assessor idiosyncrasy or between-assessor variability in assessment of communication skills. Hofstede's theory has been criticized for equating culture with national identity;[14] however for the purposes of this study we are focusing on relating the cultural dimensions to individuals and not national populations as a whole. This strategy is supported by a previous study measuring cultural orientations at the individual level.[15]

Research purposes are translated into two specific research questions:

1) How are cultural orientations associated with overall global assessment scores for 3 videos portraying different patient-practitioner interactions?

2) How are cultural orientations associated with assessors' interpretation and evaluation of affective and instrumental communication behaviours?

Methods

Study Design

This was a stimulated recall study evaluating patient-practitioner interactions, using videotaped interactions as the assessment platform. Stimulated recall was used to capture

assessors' cognitive processing during observation and evaluation of performance.[16] Verbal protocol analysis[17] was used to explore how assessors interpreted and valued affective and instrumental communication behaviours to arrive at judgments about performance.

Context of the study

The study was conducted in Doha, Qatar at the College of Pharmacy at Qatar University. Qatar is a small affluent country bordering Saudi Arabia and the Arabic Gulf. The population is diverse, with local Qatari comprising only approximately 20% of the people residing in Qatar.[18] Expatriates comprise the majority of the population and come from all world regions, primarily South Asia, Philippines, Middle East, North Africa, and Western countries. Healthcare and education sectors display similar ethnic diversity of working professionals. As part of a national vision, these sectors are undergoing major reforms to be in line with North American and European models and standards.[19] The setting is representative of a region with high ethnic diversity.

Participants

We enrolled practicing pharmacists and pharmacy educators with faculty or clinical appointments at the College of Pharmacy at Qatar University. Subjects were eligible for this study if they had experience assessing communication skills of pharmacists or pharmacy students during experiential training internships or campus-based objective structured clinical examinations (OSCEs). Convenience sampling was used by sending personalized emails to potential subjects that provided study objectives and procedures and sought interest in participation. If interest to participate was expressed, subjects were given more details regarding the study and were required to provide informed consent. Recruitment took place over 2 weeks until the target sample size of 25 subjects was obtained. This target was based on previous exploratory studies in qualitative research employing similar methods and analysis.[20,21]

Pilot

A small pilot was completed prior to enrolling subjects for the larger study. The purpose of the pilot was to select videotaped patient-pharmacist interactions for assessment in the larger study, to refine study methods and interviewing techniques prior to study implementation. Three expert- assessors were purposively chosen as experienced communication evaluators, in order to provide accurate baseline assessments of communication behaviours as displayed in the videos and to determine performance levels of videotaped candidates. Three pilot videos were chosen for inclusion in the study. The only procedural change based on pilot data was to formalize stop points during the stimulated recall procedure every 30 seconds to ensure consistency between subjects.

Measurement

Assessors' cultural orientations were measured through use of CVSCALE (Individual Cultural Values Scale). CVSCALE is a validated questionnaire to measure cultural orientations on an individual level.[15] The scale consists of 26 statements that subjects rank on a 5-point Likert scale of agreement (power-distance, uncertainty avoidance, masculinity, individualism) or importance (long-term orientation). We excluded items for long-term orientation due to no known influence on affective or instrumental communication behaviours, as per Meeuwesen et al.[11]

Three videos served as the patient-practitioner interactions to be evaluated.[22-24] The duration of each video was between two and five minutes and each video documented an interaction between a patient and a pharmacist. Videos were carefully selected based on differing performance levels of the practitioner for both instrumental and affective communication behaviours as determined from performance level assessments and global rankings during the pilot. Video scenarios are summarized in Table 2. The study was capped at three videos, in order to ensure procedure time remained feasible for recruited participants (up to 1 hour total).

Table 2. Descriptions of selected videos

	Video 1	Video 2	Video 3
Setting	Outpatient	Outpatient	Outpatient
Pharmacist description	Female pharmacist (30-40 years old)	Male pharmacist (30-40 years old)	Female pharmacist (20-30 years old)
Patient description	Elderly male patient	Angry mother of a 15 year old daughter	Female (20-30 year old) patient
Interaction type	Warfarin counselling	Response to mother's concerns regarding daughter's contraceptive	Counselling on a cholesterol-lowering medication
Details	The pharmacist gives a brief overview of warfarin, although she generally mentions most counselling points. She appears nervous and does not react to the patient's verbal and nonverbal communication cues. However, she asks the patient good questions and provides opportunity for follow up.	The pharmacist responds to the angry mother by maintaining his composure and providing her with psychosocial advice regarding interacting with her daughter. He also acts within the laws and regulations of the Canadian context by not breaching patient confidentiality. The mother is visibly satisfied at the end of the interaction.	The pharmacist gives a poor performance with respect to gathering and providing information. She fails to clarify unclear points and consistently makes assumptions regarding the patient's medical and medication history. She is friendly and personable (uses social talk) yet her tone is described as 'unprofessional' and occasionally 'harsh'. She summarizes the interaction but globally fails to engage the patient during the interaction.
Pharmacists' Instrumental Communication	**Positive – Mixed** (asks good questions and gives information but counselling lacks depth)	**Positive** (provides excellent information and rationale)	**Negative** (fails to ask questions and provide information to the patient)
Pharmacists' Affective Communication	**Negative** (pharmacist deemed unconfident, demonstrated poor eye contact, and fails to engage patient)	**Positive** (demonstrated empathy, uses good nonverbal communication, shows understanding, but at	**Mixed** (friendly and personable yet occasionally gives harsh and unprofessional tone)

		times harsh in tone and word choice)	

A 5-point Likert global assessment scale was used to capture assessors' judgements regarding pharmacist performance in each video. The 5-point scale consists of 3 descriptors that focused on global performance. Assessors were allowed to give scores from 1 to 5 at 0.5 increments.

Procedures

Step 1:

Subjects completed the CVSCALE prior to meeting with the investigators.

Step 2:

After a brief orientation to the process and assessment instruments, subjects were shown three videos separately in a random order. Subjects were told to focus assessments on 'communication skills' and 'global performance' and were allowed to take notes throughout. After each video finished, subjects were asked to rank the performance of the pharmacist according to the global assessment tool described above.

Step 3:

After initial ranking of global performance, participants were asked to complete a 30-60 min stimulated recall interview. The same interviewer (KW) completed all interviews to ensure standardization. A second person was present in 13 of 25 interviews to monitor protocol methods and ensure validity of data obtained. Immediately following rating of performance, the voice recorder was turned on and subjects were asked to verbally explain their rationale for the ranking. Next, subjects re-watched the video while being voice recorded and were instructed to verbalize thoughts regarding the pharmacist's performance at any point during the interaction. If nothing had been said for thirty seconds, the researcher stopped the video and asked if subjects had any comments from the previous segment. The same process was repeated for all three videos.

Data Analysis

Item scores for each dimension on CVSCALE were averaged to determine subject scores on each cultural dimension.

Immediately following interviews, transcripts of recordings were produced verbatim and were validated by a research assistant. Once all transcripts were produced, they were segmented into phrases by one investigator (KW) with each segment representing a single thought or idea. Then, each segment was assigned a coding category based on a pre-defined coding framework, developed from the study by Meeuwesen et al.[11] and confirmed using known instrumental and affective behaviours.[12] Table 3 presents the final coding framework and categories.

Table 3. Final coding framework per categories of instrumental and affective communication behaviours

Category	Codes
Instrumental	Identified problem Listens attentively

	Confirms problem and screens further Negotiates agenda Elicits history Uses open and closed questioning Facilitates responses Periodically summarizes Uses concise questions and comments Avoids jargon Establishes timeline Systematic/organized Summarizes to confirm understanding Transitions Logical sequencing Appropriate timing Shares thinking Explains rationale Chunks and checks Assesses starting point Asks pertinent questions Gives information Adheres to regulations Clarifies information Contracts with patient for next step Provides safety nets Checks understanding Summarizes session
Affective	Greets patient Introduces self Confirms identity Obtains consent Provides privacy Demonstrates respect and interest Positive demeanor Non-judgmental Empathetic Sensitive Professional Listens attentively without interrupting Reacts to verbal and non-verbal cues Encourages expression of feelings Eye contact Facial expressions Posture, movement Gestures Rate of speech Volume of speech Tone of speech Hesitation Confidence Assertiveness Maintains composure

	Friendly/Social talk

One investigator (KW) and one research assistant independently coded 10 transcripts. Coding was compared after each transcript and any discrepancies were resolved through discussion. At this point, it was found coding matched for >90% of each transcript and one investigator (KW) completed coding for the remaining 15 transcripts. For each participant, the numbers of statements per coded category (instrumental or affective behaviours) were calculated and transformed to percentages in order to correct for between-subject variance in verbosity and elaboration of answers.

Statistical Analysis

Descriptive statistics were used to summarize subject demographics, video scores, and category frequencies. To answer our first research question, we used the nonparametric Spearman's Rank Correlation Coefficient to identify correlations between each cultural dimension score and global assessment scores per video. To answer our second research question, we used Spearman's Rank Correlation Coefficient to determine correlations between cultural dimension scores and positive, negative, and total utterances of instrumental and affective behaviour per video. For all statistical analyses, we tested the null hypothesis of no correlation (rho = 0). Statistical significance was pre-defined at an alpha level of 0.05. All analyses were completed using SPSS Statistics v.22.

Results

A total of 25 subjects (60% male) were recruited and completed the interview process for the full study post-pilot. All participants were pharmacists and had at least 3 years of practice experience. All (25/25) had experience assessing students in practice and 16/25 (64%) had experience assessing students in OSCE settings. All subjects were employed in Qatar (academic or practice settings) during the study period. Subjects came from fourteen countries of origin, which included Canada, Egypt, Fiji, Ghana, India, Jordan, Lebanon, Nigeria, Peru, Qatar, Somalia, Sudan, Syria, and the USA. Measured cultural orientations are given in Table 4. Overall assessors scored low on power-distance, high on uncertainty avoidance, and average on masculinity-femininity and individual-collectivism. The greatest variability between assessors was noted for masculinity-femininity.

Table 4: Cultural dimension scores as measured by CVSCALE

Dimension	Median (scale 1-5) (range)
Power-Distance (n=25)	1.60 (1.0-3.2)
Uncertainty Avoidance (n=25)	4.00 (2.80-4.80)
Masculinity-Femininity (n=25)	2.25 (1.25-4.25)
Individualism-Collectivism (n=25)	3.33 (2.33-4.33)

Table 5 presents correlations between assessors' cultural orientations and global score ratings. Participants gave video 2 the highest global score (good instrumental and affective), followed by video 1 (good instrumental, poor affective), and lastly video 3 (poor instrumental, mixed affective). The cultural orientations of power-distance and masculinity-femininity were significantly associated with global assessments of video 3. Specifically,

those scoring higher on power-distance and higher on masculinity gave higher scores for this video (r = .445 and r = .537, respectively). No other significant correlations were found between video scores and cultural orientations.

Table 5: Overall global assessment scores given by assessors and correlations with measured cultural dimensions

Video	Median (scale 1-5) (range)	Correlation with Power-Distance (r)	Correlation with Masculinity-Femininity (r)	Correlation with Uncertainty Avoidance (r)	Correlation with Individualism-Collectivism (r)
1 (n=25)	3.0 (2.0-4.5)	-.091 (p=0.664)	-.023 (p=0.915)	.150 (p=0.474)	.314 (p=0.126)
2 (n=25)	5.0 (3.5-5.0)	.064 (p=0.761)	-.055 (p=794)	.128 (p=0.543)	-.092 (p=0.660)
3 (n=25)	2.5 (1.0-5.0)	.445* (p=0.026)	.537* (p=0.006)	.116 (p=0.579)	.212 (p=0.308)

*denotes statistical significance ($p < 0.05$)
r = Spearman's Rank Correlation Coefficient

Table 6 presents correlations between assessors' cultural orientations and utterance proportions of positive, negative, and total instrumental and affective communication behaviours. Table 6 shows that assessor interpretations and valuing of observed behaviours in videos 1 and 2 reflect practitioner performance on the videos and does not seem to be affected by assessors' cultural orientations. Only one significant correlation was noted in video 1 (Table 6). Masculine assessors were associated with giving less positive utterances of affective behaviours (r = -.396). However, these utterances only accounted for 1.5% of total assessor utterances on this video, which precludes interpretation of this result. No associations were significant in video 2.

Table 6. Overall percentages of utterances per category and correlations between utterance proportions and cultural dimensions of assessors

Variable	Overall percentage of utterances per category (SD)	Correlation with Power-Distance (r)	Correlation with Masculinity-Femininity (r)	Correlation with Uncertainty Avoidance (r)	Correlation with Individualism -Collectivism (r)
Video 1 Instrumental					
Positive	9.56% (0.07)	-.236	-.025	.135	-.064
Negative	24.1% (0.18)	.167	-.015	-.006	-.115
Total Instrumental	33.7% (0.17)	-.100	-.066	.028	-.207
Video 1 Affective					
Positive	1.54% (0.04)	-.360	-.396*	.127	-.032
Negative	64.8% (0.17)	-.067	.120	-.067	.204

Total Affective	66.3% (0.18)	-.100	.066	-.028	.207
Video 2 Instrumental					
Positive	33.7% (0.18)	.063	.024	.056	-.022
Negative	5.74% (0.12)	.150	.078	-.226	-.017
Total Instrumental	39.4% (0.17)	.195	.199	-.031	.064
Video 2 Affective					
Positive	49.9% (0.17)	.030	-.027	-.267	-.126
Negative	10.2% (0.15)	-.066	-.307	.219	.189
Total Affective	60.2% (0.18)	-.194	-.191	.044	-.050
Video 3 Instrumental					
Positive utterances	7.62% (0.16)	.372	.363	.117	.070
Negative utterances	39.9% (0.20)	-.533*	-.529 *	.179	-.243
Total Instrumental	47.5% (0.19)	-.247	-.343	.325	-.081
Video 3 Affective					
Positive utterances	10.1% (0.13)	.354	.441*	-.062	.121
Negative utterances	42.4% (0.20)	.104	.117	-.205	.015
Total Affective	52.5% (0.19)	.247	.343	-.325	.081

*denotes statistical significance ($p<0.05$)
r = Spearman's Rank Correlation Coefficient

Table 6 also shows that assessor interpretations and valuing of observed behaviours in video 3 also generally reflected practitioner performance (Table 2). Table 6 furthermore shows that, for video 3, assessors high on power distance and masculinity gave less negative utterances of instrumental behaviours (r = -.533 and -.529, respectively). Conversely, assessors high on masculinity were associated with giving more positive utterances of affective behaviours displayed in this particular video (r = .441).

Discussion

This stimulated recall study with verbal protocol analysis attempted to answer two research questions pertaining to the effect of an assessor's cultural orientations on assessment of communication behaviours. We found that assessors scoring high on power-distance and masculinity provided fewer negative utterances regarding instrumental behaviours, resulting in higher overall scores on an interaction deemed to demonstrate borderline performance (poor instrumental and mixed affective behaviours). These findings support the notion that cultural orientations can influence assessment of communication behaviours and have several implications for assessment in settings of high cultural diversity.

We found the influence of culture on assessment is likely interaction and context or scenario dependent, and more prevalent within borderline performances. This was demonstrated through significant correlations for global scores and assessor cognitive processes on one of three video interactions (video 3) that demonstrated borderline performance, as compared to the other two interactions or videos (videos 1 and 2). Video 2 depicted a practitioner performing very well on both instrumental and affective communication behaviours, while video 1 depicted a practitioner with good instrumental but poor affective behaviours. Assessors were more likely to agree on these videos, probably due to the explicit nature of the performance levels. However, the performance levels on video 3 were less clear and therefore could be more open to discrepancies in assessor judgements. As such, cultural orientations may have had a greater role in influencing both scoring and associated judgments of these behaviours.

The results obtained from video 3 suggest that assessors higher on power-distance and masculinity may prefer closed communication behaviours during patient-practitioner interactions. As described in Table 1, this interaction portrayed a young female pharmacist interacting with a young adult female patient regarding a new cholesterol-lowering medication. The pharmacist barely provided any information regarding the medication and its associated benefits and risks. Furthermore, the pharmacist did not engage the patient, due to which the communication was not patient-centered, and primarily used closed-ended questions. As open communication and patient-centered care are purported as preferred practice in Western settings,[25] it is interesting that countries known to be high on power-distance are largely non-Western in nature (Arab, Asian, and Central American states).[9] It is possible, therefore, that open, patient-centered communication styles may be more difficult for assessors originating and trained in these countries to interpret according to Western standards.

One important finding of this study is that assessors appeared to value communication behaviours that match their own communication preferences, and these differed amongst assessors themselves within the multicultural setting of this study. In particular, the findings on video 3 were in line with the results of Meeuwesen et al.[11] in the sense that assessors high on power-distance potentially rewarded social talk, as exemplified on video 3 (Table 2). Similarly, more masculine assessors in our study rewarded affective (socio-emotional) behaviours on video 3 positively. Our findings with respect to assessor differences are in line with assessor cognition literature, suggesting that assessors hold differing perspectives and each bring varying yet valid interpretations of candidate performance.[26] Obviously, if assessment is to remain a responsibility of a trained professional observing the interaction, it can be argued that multiple assessors with varying cultural backgrounds may be a preferred strategy to capture judgments from differing perspectives. More importantly, however, our findings illustrate the need to include teachers and trainers from various cultural backgrounds in training of communication skills, to promote student learning, to ensure that students receive feedback from various perspectives to develop adaptive communication behaviours and are well-prepared for practice in multicultural settings. Overall, our findings confirm that the perspective of the communication receiver is crucial in assessment of communication skills. This immediately raises the question of who is best suited to assess communication behaviours. In the end, it is the patient who ultimately

receives communication and can deem whether or not it was effective for them. Solely relying on assessor interpretations, rather than the communication receiver's (patient) interpretation, is not appropriate if assessors are not able to understand how the communication is influencing the receiver's experience. Findings from our study therefore emphasize the key role of the patient in assessment of practitioner's competence.

The results of our study thus have several implications for assessment of communication within patient-practitioner interactions, as well as for future research. Findings of this study reflect variations in the way performance can be understood, experienced, and interpreted in multicultural settings. Being aware of different interpretations of communication behaviours can provide valuable feedback for learners and help to better determine pass-fail decisions in both low and high stakes assessment. As indicated above, this suggests there may be a need in these settings to use multiple assessors with differing backgrounds to gain greater perspective on the communication behaviours exhibited. It also suggests that patients could be engaged in communication assessment, as they are the ultimate receiver of communication and judgments may provide more in depth feedback than what observing assessors can provide.

One may argue that assessor-training programs could be implemented to better 'standardize' assessments and provide a frame of reference according to communication norms deemed best practices within the local context. Assessor training programs have been previously implemented in attempts to increase the reliability of assessments by decreasing inter-assessor variability.[26] The problem with this approach, however, is that it does not account for these differing perspectives that could in fact be valid interpretations of performance. Based on our findings we suggest, therefore, assessor training should address cultural considerations from the patient perspective, in order to increase assessor cultural awareness and improve cultural sensitivity in assessment practices. Assessors must develop this expertise in order to adequately assess student adaptability and effectiveness of student communication behaviours within and across various contexts.

Our findings furthermore demonstrate a need to review assessment instruments for communication skills in multicultural settings. Rather than using rating scales to quantify student performance, it may be more beneficial to use qualitative or mixed methods techniques such as (standardized) narratives to gain a deeper understanding of student abilities and how communication behaviours are interpreted during assessment. This approach could also promote provision of rich feedback when used in a formative manner. Additionally, if multiple assessors are used or patients are engaged in the assessment process, this technique will allow students to have greater understanding of their performance from multiple, differing perspectives. Credibility of this approach, especially for determination of pass-fail decisions, should be evaluated in future studies.

The results of this study support future research initiatives in the area of culture and assessment. Our results demonstrate potential cultural influences on assessment, especially for students of borderline performance. Introduction of high stakes performance-based assessments in settings of great cultural diversity may be prone to discrepancies in pass-fail decisions and future studies should explore assessment methods and tools that can ensure defensible and robust decision making in these complex assessment settings. The

effectiveness of the strategies we have outlined, including the use of multiple assessors and qualitative assessment techniques, should be an immediate research priority. Finally, the association of cultural orientations with both scores and cognitive processes generates a hypothesis that provision of feedback following direct observation of communication behaviours likely differs based on cultural background of assessors. This finding justifies further study on the impact of culture on formative assessment, feedback and use of feedback in development of communication skills.

Our findings must be interpreted in light of the limitations of our study. First, our sample size of assessors was small (n=25). This likely limited our power to detect significant correlations with cultural dimensions, which may have led to type II errors. The small sample size also resulted in minimal distribution across some of the cultural dimensions, which again may have resulted in a homogenous sample and limited the ability to detect meaningful correlations outside of power-distance and masculinity. Secondly, only three patient-practitioner interactions were evaluated. As no previous information was available to guide choice of interactions, we decided to choose based on performance levels with respect to instrumental and affective communication behaviours, identified by expert-assessors. Based on findings from our study, however, future studies should choose interactions of borderline performance, as these interactions are likely more prone to variations explained by differing cultural beliefs. Finally, this study evaluated the broad categories of instrumental and affective communication. There are many specific communication behaviours, though, that may be interpreted and assessed differently within various cultural contexts (eye contact, confidentiality, physical touch) and these warrant further investigation. However, in light of these limitations, this study is the first to explore the effect of cultural orientations on assessment of communication behaviours. It generates meaningful practical implications for assessment settings high in cultural diversity, as well as suggestions for further research on effects of increasing globalization on health care professional training and assessment.

Conclusions

The results of our exploratory study give greater understanding to the effect of culture on assessment of communication skills. Assessors' cultural dimensions of power-distance and masculinity may influence assessor ratings, interpretations, and evaluations regarding communication behaviours in patient-practitioner interactions of borderline performance levels. This finding may have major implications on definitions of 'accurate' or 'correct' communication behaviours, student pass-fail decisions, and learner feedback. Contrary to current assessment practices of standardization and assessor calibration, these results support the use of multiple assessors with varying cultural backgrounds in multicultural settings, as well as further investigation into the use of qualitative or mixed quantitative-qualitative assessment methods. The results also warrant investigation into whether or not the communication receiver (i.e. patient) should be involved in the assessment process. Future studies should address these considerations, in order to better understand the effects of culture on assessment of communication skills, the effectiveness of assessment practices within multicultural settings, and any impact of assessment on patient care and health outcomes.

References

1. Ha JG, Longnecker N. Doctor-patient communication: A review. Ochsner J. 2010;10:38-43.

2. Simpson M, Buckman R, Stewart M, Maguire P, Lipkin M, Novack D, Till J. Doctor-patient communication: the Toronto consensus statement. Brit Med J. 1991;303:1385-7.

3. Merriam-Webster. Simple definition of culture. http://www.merriam-webster.com/dictionary/culture/. Accessed 14 Sep 2016.

4. Kroeber, AL, Kluckhohn C. Culture: a critical review of concepts and definitions. Cambridge: Harvard University Press; 1952.

5. Thomas SB, Fine MJ, Ibrahim SA. Health disparities: the importance of culture and health communication. Am J Public Health. 2004;94:2050.

6. Beach MC, Price EG, Gary TL, Robinson KA, Gozu A, Palacio A, et al. Cultural competency: a systematic review of health care provider educational interventions. Med. Care. 2005;43:356-73.

7. Kumagai AK, Lypson ML. Beyond cultural competence: critical consciousness, social justice, and multicultural education. Acad Med. 2009;84:782-7.

8. van Widenfelt BM, Treffers PD, de Beurs E, Siebelink BM, Koudijs E. Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. Clin Child Fam Psychol Rev. 2005;8:135-47.

9. Hofstede G. National cultures in four dimensions: a research-based theory of cultural differences among nations. Int. Studies Management & Organization 1983;13:46-74.

10. Hofstede G. Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. 2nd ed. Thousand Oaks, CA: Sage; 2001.

11. Meeuwesen L, van den Brink-Muinen A, Hofstede G. Can dimensions of national culture predict cross-national differences in medical communication? Pat Educ Couns. 2009;75:58-66.

12. Sandvik M, Eide H, Lind M, Graugaard PK, Torper J, Finset A. Analyzing medical dialogues: strength and weakness of Roter's interaction analysis system (RIAS). Pat Educ Counsel. 2002;46:235-41.

13. Laroche L, Rutherford D. Cross-cultural communication. In: Laroche L, Rutherford D, editors. Recruiting, retaining and promoting culturally different employees. New York: Routledge, Taylor & Francis Inc.; 2007. p. 99-162.

14. Baskerville RF. Hofstede never studied culture. Account Organ and Soc. 2003;38:1-14.

15. Yoo B, Donthu N, Lenartowicz T. Measuring Hofstede's five dimensions of cultural values at the individual level: development and validation of CVSCALE. J Int Consum Market 2012;23:193-210.

16. Calderhead J. Stimulated recall: a method for research on teaching. Br J Educ Psychol. 1981;51:211-7.

17. Chi M. Quantifying qualitative analyses of verbal data: a practical guide. J Learn Sci. 1997;6:271-315.

18. Qatar Population Statistics 2012: Three years after launching the population policy. http://www.gsdp.gov.qa/portal/page/portal/ppc/PPC_home/ppc_news/ppc_files_upload/populations_status_2012_en.pdf/. Accessed 14 Sep 2016.

19. Qatar National Vision 2030. http://www.mdps.gov.qa/portal/page/portal/gsdp_en/qatar_national_vision/qnv_2030_document/QNV2030_English_v2.pdf/. Accessed 14 Sep 2016.

20. Govaerts MJB, Van de Wiel MJW, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ. 2013;18:375-96.

21. Naumann FL, Marshall S, Shulruf B, Jones PD. Exploring examiner judgement of professional competence in rater based assessment. Adv in Health Sci Educ. 2016 DOI:10.1007/s10459-016-9665-x

22. Standardized patient sample interview. https://www.youtube.com/watch?v=A-XIWn1UfKg/. Accessed 14 Sep 2016.

23. PEBC Pharmacist OSCE Practice Station – Scenario by Pharmacy Prep. https://www.youtube.com/watch?v=y1WFiTrNvfg/. Accessed 14 Sep 2016.

24. Bad Patient Counseling Example. https://www.youtube.com/watch?v=OHiovB-JuJw/. Accessed 14 Sep 2016.

25. Davis K, Schoenbaum SC, Audet A. A 2020 vision of patient-centered primary care. J Gen Int Med. 2005l;20:953-7.

26. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. Med Educ. 2014;48:1055-68.

Chapter 3

Discriminating features of narrative evaluations of communication skills during an OSCE

Published as: Wilby KJ, Govaerts MJB, Austin Z, Dolmans DHJM. Discriminating features of narrative evaluations of communication skills during an OSCE. Teach Learn Med 2019 (early online). DOI:10.1080/10401334.2018.1529570

Abstract

Construct
Authors examined the use of narrative comments for evaluation of student communications skills in a standardized, summative assessment (OSCE).

Background
The use of narrative evaluations in workplace settings is gaining credibility as an assessment tool but it is unknown how assessors convey judgements using narratives in high-stakes standardized assessments. The aim of this study was to explore constructs (i.e. performance dimensions), as well as linguistic strategies that assessors use to distinguish between poor and good students when writing narrative assessment comments of communication skills during an OSCE.

Approach
Eighteen assessors from Qatar University were recruited to write narrative assessment comments of communication skills for 14 students completing a summative OSCE. Assessors scored overall communication performance on a 5-point scale. Narrative evaluations for the top and bottom two performing students for each station (based on communication scores) were analyzed for linguistic strategies and constructs that informed assessment decisions.

Results
Seventy-two narrative evaluations with a total of 662 comments were analyzed. Most comments (77%) were written without the use of politeness strategies. A further 22% of comments were hedged. Hedging was used more commonly in poor performers, compared to good performers (30% vs. 15%, respectively). Overarching constructs of confidence, adaptability, patient safety, and professionalism were key dimensions that characterized the narrative evaluations of students' performance.

Conclusions
Results contribute to our understanding regarding the utility of narrative comments for summative assessment of communication skills. Assessors' comments could be characterized by the constructs of confidence, adaptability, patient safety, and professionalism when distinguishing between levels of student performance. Findings support the notion that judgements are arrived at by clustering sets of behaviors into overarching and meaningful constructs, rather than by solely focusing on discrete behaviors. These results call for the development of better-anchored evaluation tools for communication assessment during OSCEs, constructively aligned with assessors' map of the reality of professional practice.

Introduction

There are increasing calls for use of narrative evaluations for assessment within medical education.[1,2] Narrative allows assessment to go beyond quantitative scales and rubrics and enables greater individualization and specificity of assessment information.[3-5] These data can be used to inform targeted feedback in formative assessment and to support decisions in summative assessments that must be defendable (such as graduation decisions, or licensure and registration requirements).[3,5,6] Moreover, narrative data may allow decision-makers to better conceptualize student performance, based on better understanding of assessors' performance interpretations and judgements.

Although, theoretically, utility of narrative comments for assessment purposes is established, questions remain prior to implementation as a summative assessment strategy.[6] Past research on narrative from workplace-based in-training evaluation reports (ITERs), for example, showed assessors can reliably categorize and rank order narrative evaluations based on trainee overall performance level and clinical skills.[5,6] Interestingly, however, little research has tried to determine how those assessors who write narrative use certain words, comments, or phrases to provide this discriminatory ability, especially relating to communication competencies. For example, it is unknown which communication skills, behaviors, or attributes assessors document in support of high performance and vice versa for justification of poor performance. It is known from rater cognition research that assessors typically vary on what they focus on when observing interactions and how they interpret overall performance when making judgements as a whole.[7-9] There is growing evidence that these differences could represent meaningfully idiosyncratic interpretations of student performance.[8] Consequently, in narrative comments assessors may be inclined to document components that are relevant to their own performance theory, namely the unique set of behaviors or constructs that they pay attention to when making an overall judgement.[4,10-11]

Aside from *what* assessors document to support or justify high or poor performance, *how* they write the comments may offer additional insight into their overall performance judgment. Research findings in the context of workplace-based assessments indicate that assessors' language 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.[12-14] Findings demonstrate that assessors frequently use linguistic strategies of politeness to disguise or soften conveyed judgements, more commonly in poor performers.[14] Politeness theory attempts to explain how individuals in the setting of a face-threatening act use linguistic strategies protect the receiver's self-image as well as to evade any potential conflicts and ensure smooth social interactions.[15] With respect to poor performers, for example, the assessor may modify the impact of their comments by using strategies to soften effects on the receiver, such as depersonalizing comments ("no eye contact" vs. "*she* gave no eye contact") or choosing

hedges (words/phrases) to not fully commit to statements (" *a bit* rude", "*not very* self-confident").[14,15] In the case of high performers, on the other hand, assessors may choose to give compliments ("Excellent job") or make the receiver feel accepted as part of a group ("A great pharmacist"), to intensify their approval.[14] Research findings suggest that assessors modify their language when writing assessment comments depending on performance level (increase hedging in poor performers), affecting overall interpretation or actually providing clues or codes to assessors' intended judgement.[14]

Although data regarding narratives in assessment is increasing, research on the use of narratives for summative assessment so far focuses on workplace based settings. Given increasing calls for use of narrative assessments on the one hand, and the potential impact of assessors' idiosyncrasies and linguistic strategies on utility of narrative comments for decision-making on the other, there is an urgent need to explore these phenomena in different assessment contexts. Objective Structured Clinical Examinations (OSCEs) are typically task-focused assessments that are used as summative evaluations of a trainee's ability to demonstrate clinical skills.[16] Recent research regarding narrative in OSCE contexts focused on the feedback potential of narratives but did not address assessors' use of narratives for high-stakes decision-making.[17,18] If narrative evaluations are to be used as a summative assessment approach in these contexts we need a better understanding of how assessors use specific comments to discriminate and convey messages about students' performance levels. Therefore, the aim of this study was to explore constructs (i.e. performance dimensions), as well as linguistic strategies that assessors use to distinguish between poor and good students when writing narrative assessment comments of communication skills during an OSCE.

Methods

Setting

The study was conducted in Qatar. The College of Pharmacy at Qatar University maintains full accreditation for the Bachelor of Science in Pharmacy and Doctor of Pharmacy programs from the Canadian Council of Accreditation of Pharmacy Programs.[19] As such, teaching and assessment methods, including use of OSCEs, align with Canadian policies, procedures, and standards.[20] Currently, students take a summative OSCE prior to graduation as an exit-from-degree exam. The OSCE is high stakes in nature, as it serves as the final examination for the final clinical course within the curriculum. For the OSCE, cases are blueprinted to the program's competency framework.[21] Interactive (i.e. communication) stations include cases related to medication counselling, referral, self-care, device teaching, ethics in confidentiality, adverse effect management, and provision of health-related information. Students are unaware of case topics prior to completion of the OSCE and have eight minutes to complete each station. Standardized textbook references are available, if required.

Research procedures and data collection

For this study, two OSCE cycles (each consisting of the same 9 interactive OSCE stations) occurred on the same day with n=7 participant-students in each cycle (total of 14 students). Students were convenience sampled from a total population of 25 students (all female), as only 25 students are enrolled per academic year. Recruitment occurred via email to all eligible students. The first 14 students to reply with interest for the study were selected. Participating students were positioned between non-participant students within each cycle, in order to allow for ample time for assessors to write a narrative evaluation. All students had experience completing formative OSCEs. All received extensive communication training both on campus during professional skills courses and off campus during 960 hours of patient care and professional practice-related activities.

A total of 18 assessors (n=2 per OSCE station) were recruited via email to provide narrative evaluations of students' communication skills for study purposes. Two other assessors were present within the station to evaluate students for grading purposes (required by university policies). Participants were selected using convenience sampling from the assessor pool at Qatar University. Assessors were eligible for participation if they were health professionals, had previous training and experience assessing pharmacy student communication skills and provided consent. All assessors were pharmacists or faculty within the College of Pharmacy. Assessors were all experienced OSCE assessors and were trained according to the framework for teaching and assessing communication skills currently used at Qatar University. For this study, they received additional training during a group session in advance of the OSCE by explaining study objectives and providing samples of narrative assessment comments extracted from the literature.[10, 22] These examples were not related to communication, so as to prevent anchoring or seeding bias during the actual study.

Assessors remained constant for each station throughout both cycles and were asked to write narrative evaluations of students' communication skills during the observed interaction. Evaluations were handwritten on a blank sheet of paper. Instructions to assessors were: "Please use the space below (and on the reverse if needed) to write a detailed narrative evaluation of the students' communication skills". No direction in the length or content of assessment comments was purposefully given, in order to minimize bias in terms of the skills, behaviors, and other attributes that assessors focus on. Assessors were given 17 minutes to write narrative assessment comments per station/student. Assessors were instructed to keep the narrative comments strictly confidential from their co-assessor or any other person to avoid data contamination.

Assessors were also asked to assign an overall performance score to each student according to a 5-point communication assessment scale with anchors at 1, 3 and 5 points (1 = Communicates inappropriately and ineffectively to the task, 3 = Communicates with some logic and comprehension but not applied consistently, 5 = Communicates precisely,

logically and perceptively to the encounter, integrating all relevant components). Assessors were instructed that their scores would not count for grading purposes but students would receive their narrative comments upon completion of the summative OSCE.

As we specifically sought to explore constructs (i.e. performance dimensions) and linguistic strategies that assessors use to distinguish between poor and good students when writing narrative assessment comments, we purposively collected and analyzed the narrative evaluations for the top two and bottom two performing students for each station. For each station, poor performers (bottom two per station) and high performing students (top two per station) were identified based on the composite quantitative communication scores given by the two assessors involved in the study.

Data analysis

Use of linguistic strategies

Extracted narrative comments were coded line by line using a deductive approach according to the adapted politeness theory framework.[14,15] If a politeness strategy was noted, it was coded as such. If no politeness strategy was noted, it was coded as 'no politeness strategy'. Upon reading the narratives it was clear the majority of comments were depersonalized and evident that depersonalization was not being used to convey politeness in this context, but rather to record short observations. We therefore coded each statement as either personalized or depersonalized, in addition to any other politeness strategy present. Comments that were depersonalized but that did not include another politeness strategy were coded as 'no politeness strategy'. Coding occurred by two independent coders, the PI (KW) and a research assistant. Coding and coding processes were regularly discussed within the research team; all investigators were given the raw coding data to review and question or challenge coding results. Disagreements for coding were resolved through discussion. The proportion of narrative comments coded for each politeness strategy was calculated. The Chi-squared test was used to compare proportions of linguistic strategies used across performance levels (IBM SPSS Statistics, Version 24.0. Armonk, NY:IBM Corp.).

Use of performance dimensions and constructs

In order to determine which aspects of communication assessors focused on within narrative when discriminating between students, comments were recoded using a general inductive coding approach.[23] Specifically, each narrative was first segmented into phrases representing a single thought, idea or statement by one of the researchers (KW). Segments were identified on the basis of semantic features (i.e. content features, as opposed to non content features such as syntax); segments could thus include several comments. Each segment was reviewed and coded by two independent coders, using open and axial coding to identify communication behaviors as well as core meaning.[23] After coding was complete, coders identified segments that appeared to be decisive for

the overall judgement, justifying the respective good or poor performance rating as determined by the quantitative scores. Identification of these segments occurred by comparing and contrasting the performance score with comments that were either positive or negative, depending on the score. For example, for a narrative with a score of '5', all segments within the narrative that represented a clearly positively perceived aspect of communication were extracted for further review. For each of these segments, narrative comments were re-viewed multiple times to search for patterns in what assessors appeared to focus on when discriminating between levels of student performance. These patterns of interrelated comments were then grouped into key categories representing different constructs that appeared to form the central focus in assessors' narrative comments. Narratives were then re-read multiple times to check for disconfirming evidence and to ensure robustness of the analysis. Again, coding and coding processes were regularly discussed within the research team and all investigators agreed on final identified constructs. Finally, representative quotes from the narrative comments were extracted.

This study was approved by Qatar University Institutional Review Board (QU-IRB 571-E/16).

Results

Assessor demographics are provided in Table 1. For 8 stations, the range of communication scores was from 2 to 5 and for 1 station it was from 3 to 5. Seventy-two narrative evaluations were extracted for analysis (36 from good performers and 36 from poor performers). The final data set included a total of 662 individual comments, averaging 9 comments per narrative.

Table 1. Baseline characteristics of assessors (n=18)

Assessor Characteristic	Number (%)
Gender	
Male	8 (44.4)
Female	10 (55.6)
English Proficiency	
Native English speaker	7 (38.8)
Non-native English speaker	11 (61.1)
Job Role	
Faculty	9 (50%)
Practicing Clinicians	6 (33.3%)
Both	3 (16.7%)
Highest Degree Obtained	
Bachelor	1 (5.56%)

Master	2 (11.1%)
Doctor of Pharmacy	10 (55.6%)
Doctor of Philosophy	4 (22.2%)
Doctor Naturopathic Medicine	1 (5.56%)
Years of Experience Assessing Communication Skills	
<5	12 (66.7%)
5-10	4 (22.2%)
>10	2 (11.1%)

Use of linguistic strategies

Table 2 provides results of the linguistic line-by-line coding analysis. The majority of narrative comments were written with no politeness strategy (77%) and were depersonalized (73%). A further 22% of comments were hedged. With respect to the entire narrative, 8 (11%) of narratives contained no politeness, 17 (24%) were entirely depersonalized, and 51 of 72 (71%) contained at least one hedge. Assessors used politeness strategies differently between good and poor performers. Coding of 'no politeness strategy' was more common in comments pertaining to good vs. poor students (83% vs. 70%, respectively, chi squared test statistic 15.3, $p<0.001$). Table 2 furthermore shows that hedging was more commonly coded in comments for poor performers, as compared to good performers (30% vs. 15%, respectively, chi squared test statistic 21.4, $p<0.001$).

Table 2. Common linguistic strategies coded within narratives obtained from OSCE evaluations (n=9 stations, n=18 assessors) for good and poor performing students

Category	Definition	Example	Poor Students	Good Students	Total
Total Comments			328	334	662
No Politeness	No politeness strategies	"Used lay language"	231 (70%)[a]	278 (83%)[a]	509 (77%)
Depersonalization	Does not directly refer to receiver using names, or other pronoun (i.e. she, he, him, her)	"Good eye contact, posture and clear voice tone" "Didn't encourage questions"	230 (70%)	252 (75%)	482 (73%)
Hedging	A word or phrase used to modify the degree of membership in a set (i.e. lacking commitment to statements made)	"She was *a bit* rude when talking to the patient" "*Could be* more assertive and confident in response" "I felt she was *not very* self-confident when she was speaking to the patient"	97 (30%)[b]	49 (15%)[b]	146 (22%)

[a] chi squared test statistic 15.3, $p<0.001$; [b] chi squared test statistic 21.4, $p<0.001$

Use of performance dimensions and communication constructs

We identified four communication constructs that characterized the written assessment comments (i.e. confidence, adaptability, patient safety, and professionalism). Behaviors were documented both generally (usually for good performers) or specifically using detailed descriptions of student behaviors within the interaction (usually for poor performers). The following sections provide a description of each of the constructs and how assessor comments characterized the construct, with specific examples for both good and poor

performers. Explanations of how the comments were deemed to be coupled with assessors' communication scores are also provided.

Construct 1: Confidence

Confidence was identified as a key characteristic of assessors' narratives, often explicitly stated and exemplified using a cluster of different behavioral descriptions:

> "Projected well, was confident and assertive" (Assessor 7, Student 14, Score=4)

> "Tone of voice is confident and reassuring…No distracting hand/body gestures" (Assessor 14, Student 13, Score=5)

For poor performers, comments often included more detailed descriptions of specific behaviors, inferring (lack of) confidence:

> "When making recommendation: very soft spoken and not confident, uses words such as 'like something' not sure of self" (Assessor 3, Student 2, Score=3)

> "Backed away when patient was upset…Fiddled with pen during entire conversation…not confident with answer" (Assessor 8, Student 2, Score=2)

For both good and poor performers, these comments about confidence appeared to set the tone of the overall narrative. The common attribute between most confidence-related comments was that of demonstrating command of the interaction. As shown in the examples above, assessors used specific communication behaviors (e.g. gestures or eye contact) to explain or justify how a student did or did not portray confidence. These could be explicitly stated ("poor eye contact – lack of confidence") or implicitly interpreted based on the comments in context of the segment or complete narrative. Voice, with respect to tone and volume, was central across the entire data set as a measure of confidence, clearly discriminating between good and poor performers.

Construct 2: Adaptability

Similarly, the ability of students to adapt or tailor communication to the patient was identified as a central construct in assessor narratives. Good performers were explicitly rewarded for being patient-centered in their approach, exemplified by verbal and nonverbal behaviors (voice tone, facial expressions, and gestures/body language) and the patient's satisfaction with the interaction, as inferred from verbal or non-verbal patient communication:

> "Expresses compassion, responds perceptively to patient feelings…Establishes good rapport with the patient" (Assessor 18, Student 10, Score=5)

> "The student showed excellent empathy towards the patient, and she used good non-verbal gestures to make the patient more comfortable even if she was in a hurry" (Assessor 1, Student 10, Score=5)

> "Stood to match patient, asked him to sit, good to establish rapport" (Assessor 7, Student 14, Score = 4)

Conversely in poor performers, assessors documented inability to develop rapport, and inappropriate responses to the patient's reactions and emotions:

> "Smiled throughout interaction despite patient being upset…Did not acknowledge patient's emotions…Did not try to explain answer in any other way" (Assessor 8, Student 2, Score=2)

> "Slightly superior mannerisms…Not listening to standardized patient's intro – in too much of a hurry…Not reassuring – no warmth in manner. This mother is worried – it's your job to calm her down" (Assessor 14, Student 4, Score=3)

> "She did not show sympathy to the patient and talked like robot!" (Assessor 9, Student 5, Score=3)

A common feature of these comments is the incorporation of multiple examples or behaviors into a segment that, when read as a whole, represents the larger construct of the student's ability (or inability) to adapt to the situation at hand.

Construct 3: Patient Safety

Narrative comments were characterized by single statements relating to a patient's health and well-being. Safety was not always associated with a specific risk or harm done, but also included communication to promote patient understanding of the therapeutic plan and medication-related instructions. Instances where safety was mentioned positively were associated with good performance:

> "She…cared for the patient's safety regardless of the fact that he was in a hurry" (Assessor 1, Student 5, Score=5)

> "She was empathetic and noticed shortness of breath quickly and asked about it…She insisted on referral (very good)" (Assessor 5, Student 12, Score=4)

Explicitly mentioning risks to patient safety was associated with poor performance:

> "Regardless of the fact that she provided the right recommendation, she had enough time to counsel the patient about what she missed in order to ensure safety" (Assessor 1, Student 3, Score=2)

> "Referred to another pharmacist – dangerous! Did not act with responsibility and tried to pass off to another pharmacist or for patient to look it up [themselves]" (Assessor 8, Student 2, Score=2)

Although the construct of safety was identified less frequently than confidence or adaptability, the association with quantitative scores was very strong. A single mention of

the safety construct appeared to be decisive for the assessor's overall judgement of the interaction.

Construct 4: Professionalism

Professionalism was identified to be a fourth key construct in assessor narratives and comments that helped distinguish between student levels of performance. Assessors used examples relating to attitude, appearance, and professional identity to infer students' professionalism. Comments for good performers used frequently strong compliments and punctuation to accentuate emphasis:

> "For me, she demonstrates a model healthcare provider!" (Assessor 5, Student 6, Score=5)

> "Good introduction, looks professionally dressed…Overall, she discussed the situation well and had excellent verbal and nonverbal communication skills" (Assessor 7, Student 4, Score=5)

Comments for poor performers were consistently justified or explained as to why the assessor deemed the student's actions to not be professional:

> "Really needs to work on professional appearance like hair style and clothing – part of public confidence" (Assessor 14, Student 2, Score=3)

> "Not friendly – she introduced herself but not in the friendly manner" (Assessor 9, Student 8, Score=2)

For this construct, assessors appeared to distinguish between good and poor performers by providing insight of how they related to students on a professional level. Assessors focused on if the student behaves, dresses, or overall acts like an individual fit-for-practice and could be entrusted to start working as 'one of them'.

Discussion

This study attempted to explore assessors' use of narrative assessment comments of communication skills to distinguish between performance levels during a summative OSCE. We found that assessors' narrative primarily included depersonalized comments without the use of politeness yet hedging was used more commonly for poor performers. Our findings also showed that assessors' comments could be characterized by four constructs related to communication, notably confidence, adaptability, patient safety, and professionalism, when discriminating between students. Despite documenting judgements regarding many aspects of communication, including specific behaviors, it was these four constructs that consistently appeared to inform overall performance judgments as demonstrated by scoring as a 'good' or 'poor' performer.

A key finding from this study is that assessors seem to focus on fundamental constructs rather than discrete behaviors, when judging task performance and conveying performance information during OSCEs. Assessors tend to cluster specific communication behaviors into

meaningful patterns in order to explain or justify their judgements pertaining to overarching constructs. They are likely less worried if a student makes a behavioral mistake and more concerned if the student can adopt and demonstrate these core constructs or fundamental patterns of behavior that determine effective pharmacist-patient communication. This finding is in line with recent literature on rater cognition that suggests that assessors, although valuing different behaviors in assessment, tend to prioritize centralized themes when making judgements.[7,9-11,13,24] Our results support this concept and provide evidence that assessors, when writing narratives, seem to take a more holistic approach to assessment of OSCE performance, focusing on broad constructs, yet substantiate their judgements by providing descriptions of specific behaviors. Findings from our study thus advance our understanding of how assessors' observe and judge students' communication skills in standardized assessment tasks. Results from our study specifically point at key dimensions in assessors' performance theories (i.e. the types of constructs that assessors focus on when interpreting observed communication behaviors), underpinning their decisions and performance feedback.

Although contexts were not directly compared in this study, our results with respect to linguistic strategies in the OSCE setting differ from previous findings in workplace-based assessment.[14] In workplace settings, research findings showed high frequency use of positive politeness strategies (compliments, exaggeration, in group identity markers, offers, and optimism) whereas these strategies were negligible in our study.[14] As well, hedging appeared to be more commonly used by assessors in the workplace context (hedging present in 94% vs. 71% of low vs. high performer comments, respectively).[14] The specific and depersonalized feedback observed in our study may indicate that assessors feel safe to write these comments in an OSCE context. Various factors may explain this contextual difference. An OSCE station is a short, task focused interaction, compared to weeks or months of supervised, multifaceted workplace-based training. This may facilitate assessors to simply write down what they see and how they see it, rather than providing general comments, encouragement, or compliments. Additionally, as assessors within the OSCE station do not interact with students, assessors and students do not develop working relationships or interdependence. Assessors may therefore not feel the need to disguise their comments with politeness.

Implications

The two major findings of our study contribute to better understanding of the utility of narrative comments in summative assessment and have further implications for assessment of communication skills during OSCEs. The overall lack of politeness strategies within comments facilitate fairly easy interpretation of assessors' judgements, and are likely to allow students and summative decision makers to understand intended meaning of assessors' comments. Our findings with respect to assessors' use of broad constructs raise questions about common assessment practices using tools that break down communication evaluation into micro components, rather than interpreting performance as a whole. Assessment tools focusing on what assessors consider to be key constructs in communication and using assessors' language might be more constructively-aligned, more meaningful, and thus potentially more valid in high-stakes assessments.[25] Inclusion of discrete, specific behaviors, however, may be needed to enhance the assessment's formative function (specifying feedback) and/or to further justify summative judgements.

Further research for the use of narrative in summative OSCE evaluation should therefore focus on exploring the utility of anchoring summative assessment tools using overarching constructs, illustrated with specific communication behaviors. Additionally, research should focus on development of an "assessment thesaurus", listing words and phrases representing relevant performance dimensions and constructs. Development of a useful and meaningful assessment language may support assessor tasks in judgment and decision making, as well as design of rating scales with constructively-aligned anchors.

Limitations

Limitations of our study should be noted. First, our data is limited to one setting with a small sample size of students and assessors. However, while this is a limitation for transferability of the results, the number of narrative assessment comments analyzed within our study was large (72 narratives with a total of 662 comments), therefore supporting our conclusions for this specific data set. Secondly, English was not the first language for 11 of 18 assessors. Although all comments were written in English, grammar may have differed as a result of English proficiency and language and cultural differences may have influenced the linguistic analysis. Our findings, however, showed consistent use of key constructs across all assessors. A third limitation of this study is that the data were obtained from assessors who had protected time to provide comments and were not providing summative pass-fail decisions, which might explain the relative disuse of politeness strategies. It should be noted, however, that these assessors were highly trained, sampled from the same assessor pool as regular assessors, and were instructed to provide assessment as if they were making summative decisions. A final limitation of this study is that written narratives do not necessarily reflect assessors' information processing. We do not know how assessors made their judgements; i.e. if they first notice certain behaviors and thereafter construct their judgement or if they use the overarching constructs as a starting point for selecting and interpreting behavioral observations. However, providing motivations for performance ratings is inherent in many assessment tasks and our data clearly suggest that assessors' motivation in our summative communication OSCE rests on broad and overarching constructs. Despite study limitations, we feel the results are novel, meaningful and of significance for research pertaining to narrative in assessment.

Conclusions

This study contributes to increasing understanding regarding the use of narrative evaluations and was the first to explore through a qualitative lens, distinguishing features of narrative comments across good and poor performers regarding communication skills within an OSCE context. Our results support utility of narrative evaluations for summative decision making in standardized assessment settings. As evidenced from our study, narrative evaluations may enhance assessment of communication performance through facilitating development of better-anchored evaluation tools, constructively aligned with assessors' map of the reality of professional practice. In order to optimize narrative assessment approaches, however, further research and field testing across different professional and cultural contexts is needed to deepen our understanding how assessor narratives can be used to support summative decision making while fostering student learning and competence development.

References

1. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 2013;35(7):564-8.

2. 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.

3. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. Acad Med. 2016;91(10):1359-69.

4. Govaerts MJB, Van de Wiel MJW, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ. 2013;18(3):375-96.

5. 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.

6. 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.

7. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ. 2016;50(5):511-22.

8. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the black box differently: assessor cognition from three research perspectives. Med Educ. 2014;48(11):1055-68.

9. Wilby KJ, Govaerts MJB, Austin Z, Dolmans DHJM. Exploring the influence of cultural orientations on assessment of communication behaviours during patient-practitioner interactions. BMC Med Educ. 2017;17:61.

10. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. Acad Med. 2010;85(5):780-6.

11. Wilbur K, Hassaballa N, Mahmoud OS, Black EK. Describing student performance: a comparison among clinical preceptors across cultural contexts. Med Educ. 2017;51(4):411-22.

12. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "plus": the nature of written comments on internal medicine residents evaluation forms. Acad Med. 2011;86(10 Suppl):S30-4.

13. 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.

14. 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 in Health Sci Educ. 2016;21(1):175-88.

15. Brown P, Levinson SC. Politeness: Some universals in language usage. New York, NY: Cambridge University Press; 1987.

16. Austin Z, O'Byrne C, Pugsley J, Munoz LQ. Development and Validation Processes for an Objective Structured Clinical Examination (OSCE) for Entry-to-Practice Certification in Pharmacy: The Canadian Experience. Am J Pharm Educ. 2003;67:Article 76.

17. Harrison CJ, Konings KD, Molyneux A, Schuwirth LWT, Wass V, van der Vleuten CPM. Web-based feedback after summative assessment: how do students engage? Med Educ. 2013;47(7):734-44.

18. Harrison CH, Molyneux A, Blackwell S, Wass VJ. How we give personalized feedback after summative OSCEs. Med Teach. 2015;37(4):323-6.

19. Canadian Council for Accreditation of Pharmacy Programs. Accredited Programs. Canadian Council for Accreditation of Pharmacy Programs Web site. http://www.ccapp-accredit.ca. Published 2017, Accessed September 23, 2018.

20. Wilby KJ, Diab M. Key challenges for implementing a Canadian-based objective structured clinical examination (OSCE) in a Middle Eastern context. Can Med Educ J. 2016;7(3):e4-9.

21. Association of Faculties of Pharmacy of Canada. Educational Outcomes for first professional degree programs in pharmacy in Canada – June 4, 2017. Association of Faculties of Pharmacy of Canada Web site. http://afpc.info/node/39. Published June 20, 2017. Accessed September 23, 2018.

22. Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med. 2012;87(4):419-29.

23. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval. 2006;27(2):237-46.

24. Govaerts MJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. Adv Health Sci Educ. 2011;16(2):151-65.

25. Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32(3):347-64.

Chapter 4

Assessors' interpretation of narrative assessment data from a summative OSCE

Submitted

Abstract

Purpose
Increasingly, the use of narrative assessment data is being purported to enhance robustness of assessment decisions. The interpretation of written assessment comments, however, is inherently complex and relies on human (expert) judgements. The purpose of this study was to explore how assessors give meaning to narrative data when interpreting narrative assessment comments obtained from a summative OSCE.

Method
Narrative assessment comments of student communication skills as well as communication scores were obtained for 24 students across 6 stations of a summative OSCE. Aggregated narrative data across all stations was sampled for 9 students (3 good, 3 average, and 3 poor performers, based on communication score). For each of the students, ten expert-assessors reviewed the aggregated set of narrative comments. 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 assessment comments: the student (places him or herself in the shoes of the student), the examiner (adopts the role of the examiner and re-interprets comments according to their own standards or beliefs), or the professional (acts as the profession's gatekeeper by considering the assessment to be a representation of real-life practice).

Conclusion
Findings add to the understanding of assessors' interpretation of rich performance data by identifying strategies and different perspectives that expert-assessors use to frame and bring meaning to narrative comments. Assessors' perspectives affect assessors' interpretation of assessment comments, likely influenced by their beliefs, interpretations of the assessment setting, and personal performance theories. Results have implications for judgement and decision-making based on narrative assessment data and call for justification of assessor judgements and use of multiple assessors to account for variations in assessor perspectives.

Chapter 4

Introduction

There are increasing calls to capture and provide rich performance information for trainee assessment, especially for competencies such as clinical skills, communication, and professionalism.[1-3] Therefore, narrative assessment comments have been advocated for within a variety of assessment contexts, including clinical training evaluation and OSCEs.[4-6] The use of narrative comments for assessment purposes has been shown to aid judgements by capturing rich data that might be lost when using checklists, scales, or rubrics.[5,7,8] Narratives have the ability to detect student and context specific areas of strengths or weaknesses in task performance, as well as help to justify assessor judgements and decisions.[4,9] Despite narratives gaining increasing credibility for assessment and feedback, questions remain regarding their utility for decision-making.

Studies in health professionals' education have shown that assessors differ in their interpretations of rich data and reasoning of judgement.[2,10-12] Factors such as assessor expertise, assessors' beliefs and perceptions of the assessment tasks and assessment context, may result in considerable assessor variance with respect to the types of data they consider relevant for pass-fail decision making and how they use that data to reach judgements on student performance.[11-13] Differing beliefs regarding the assessment task and context, for example, may result in stricter or more lenient judgements depending on the assessors' perceptions of how the assessment relates to real-life professional practice.[14] Despite the recent developments in this area, little information exists about how assessors' approach and bring meaning to assessment comments when asked to judge performance using rich and aggregated narrative assessment data.

A study on assessors' interpretation of rich data portfolios of students in teacher education showed that large differences existed between assessors' interpretations of the same student portfolios.[15] Authors accounted for these variations in assessor interpretations by explaining that assessors, when reviewing rich data, form patterns or 'stories' of trainee performance. These stories are continually revised until all available evidence is accounted for. It is the complexity of this process and the diverse nature of the data that makes it difficult for assessors to interpret and eventually judge performance.[15] Based on the findings from their study, the authors argue that assessor variation is likely less about what assessors focus on or pay attention to when reviewing rich data and more about how they bring meaning to the data through the development of a coherent representation or story.

Clearly, before we can use narrative data for (high-stakes) decision-making, we need a better understanding of how assessors vary in their interpretations of narrative assessment comments. Given increasing calls for use of narrative comments in performance judgements, we must gain a better understanding of how assessors interpret narrative comments and conceptualize these assessment data. Understanding the underlying rationale for assessor variance may aid the design of robust assessment approaches that support the credibility of judgements and decision-making. The purpose of this study, therefore, was to explore how expert-assessors give meaning to narrative data when interpreting an aggregated set of narrative assessment comments obtained from a summative OSCE.

Methods

Study Design

This was a qualitative study 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 interpreted and gave meaning to aggregated narrative comments.

Setting

The study was conducted at the College of Pharmacy at Qatar University in Doha, Qatar. The College maintains a Bachelor of Science in Pharmacy program accredited by the Canadian Council for Accreditation of Pharmacy Programs (CCAPP). The program graduates approximately 25 female students per academic year. As part of program requirements, students complete a summative OSCE at the end of the 4-year curriculum.

Participants

A total of ten expert-assessors were recruited to review the aggregated narrative data sets 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 provision of instructions for the think-aloud procedure. No compensation was provided for participation.

Research procedures

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

For the present study, we used assessment data from the 4th year OSCE conducted in 2018. Twenty-four graduating pharmacy students completed the OSCE, which included six communication-focused stations. All stations were 8 minutes in duration. Three stations required the student to interact with a standardized patient, two with a standardized physician, and one with a standardized mother. A blueprint of the stations is provided in Table 1. For the purposes of this study, six OSCE examiners (one per communication station) were recruited to write narrative evaluations of student communication skills and to score communication skills according to a single-dimension 5-point rating scale with anchors at 1, 3 and 5 points (1 = Communicates inappropriately and ineffectively to the task, 3 = Communicates with some logic and comprehension but not applied consistently, 5 = Communicates precisely, logically and perceptively to the encounter, integrating all relevant components).[17,18] All examiners were familiar with the tool and had been previously trained using the tool via pre-assessment calibration exercises and post-assessment debriefing. Examiners also had previous training and experience 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 within a set for analysis as outlined below.

Table 1. OSCE Blueprint

Station	Standardized Actor	Description
1	Patient	A young adult female presents to a community

		pharmacy for an oral contraceptive. The pharmacist must recognize 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 recognize 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	Pediatric patient's mother	A mother is picking up an antibiotic prescription for their child. The pharmacist must provide appropriate counseling and instructions.

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) that aimed to determine the number of aggregated student narrative sets that could be realistically reviewed during study procedures without excessive burden on participants. It was determined that eight to ten sets were optimal and allowed the study protocol to be completed within 2 hours. Based on this result, we therefore stratified the 24 sets (one per student) into three groups of eight students representing low, average, and high performers based on the overall score provided by the OSCE examiners. From each of these groups we purposively selected three sets of narrative evaluations representing the top three students from the high-performance group, the bottom three students from the low-performance group, and the middle three from the average group, in order to ensure an adequate representation by performance level.

Each expert-assessor was scheduled for a meeting with the same investigator to complete a think aloud protocol.[16] Assessors were briefed on the study objectives and procedures and asked to sign written informed consent. Permission to audiotape the session was obtained. The interviewer then provided the assessor with an aggregated set of narrative evaluations (without communication scores) corresponding to one student. The order of student data sets presented to each assessor remained constant for the study, to simulate assessment approaches in real-life. The assessor was asked to begin reading the narratives as if they were evaluating the student and to verbalize all their thoughts as they emerged while reading and interpreting assessment data. The investigator prompted by saying 'please continue thinking aloud' if the assessor ceased to verbalize their thoughts for more than a few seconds. Once the assessor signalled they had finished reviewing the aggregated data for the student, the investigator ceased prompting and allowed the assessor to state any remaining impressions. The same procedure was repeated for each of the nine student narrative data sets. Field notes were taken throughout to help inform data analysis by capturing actions or behaviours that would not be analyzable from the transcript alone (e.g. facial expressions, gestures).

Data Analysis

All think aloud protocols were transcribed verbatim by one investigator (KW), who read all transcripts multiple times before coding. Transcripts were coded inductively by two independent coders (KW and a research assistant). Codes were assigned using an open-coding procedure that identified words, sentences, or phrases that related to the research question, and that reflected assessors' approaches to interpreting, giving meaning to the narratives. Coding thus occurred based on 'how' assessors brought meaning to comments, rather than the specific judgments they made. For example, a quote of "I see a pattern in comments across stations that the student had good nonverbal communication and is a good communicator" would be coded as 'comparison across stations', rather than 'good communication'. Coders compared and discussed codes throughout the coding process to clarify and resolve any discrepancies. Once all transcripts were coded, individual codes were combined into broader themes, which were discussed in the research team. Subsequently, KW and the research assistant independently conducted a between case analysis, by

comparing and contrasting patterns across cases (assessors), to search for similarities and differences in how assessors approached and interpreted narrative data. Results were discussed in the research team over multiple occasions until the final themes were agreed upon. Representative quotes were extracted from transcripts to illustrate themes and to support our interpretation of the data.

Results

We found that expert-assessors used different strategies to bring meaning to aggregated narratives, helping them to 'paint a picture', or 'build a story' of student performance. They compared and contrasted information within a student's narrative set and across narrative sets belonging to different students. Assessors also engaged in creating mental images of what happened during the OSCE, in order to better understand what occurred during the interaction. Assessors furthermore adopted different perspectives in interpretation of narrative comments, resulting in different ways of explaining, reframing or understanding written comments. The data allowed us to construct three predominant perspectives that assessors used to bring meaning to comments: the student perspective, the examiner perspective, and 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.

Strategies used in interpretation of narrative data

1. Comparing and Contrasting

For every individual student, assessors compared and contrasted comments both within and across stations in order to seek for patterns of student performance, and to arrive at a coherent interpretation of the student's communication skills. Identification and confirmation of patterns of similarity across stations facilitated development of a coherent, overall story of the student's performance:

> *(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. [A8, S3]

Once they had reviewed data from several students, assessors also began comparing between students:

> *(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]

2. 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 verbalized their images but also used hand gestures and facial expressions during the think aloud protocol to portray what was 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 patients 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 visualize 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]

Richness of the data (i.e. specific details) and presence of the examiner's interpretation of performance facilitated assessor's ability to form mental images of the student's communication behaviours. When assessors did not have enough details within the narrative to understand what occurred during the interaction (i.e. when assessors were not able to visualize), they had difficulties 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 visualize 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 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. 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 their interpreted meaning.

Student Perspective

Assessors that 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 empathized with the student and tended to relate negative behaviours to factors beyond the student's control. Assessors appeared to search

for reasons or justification for student behaviours, rather than accept them at 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]
>
> *(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 that assumed the perspective of the examiner attempted to understand the meaning of comments by placing themselves in the shoes of the examiner present. They appeared to accomplish this by assuming the role of the examiner when visualizing what occurred within the interaction. 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 interpret student performance:

> *(In response to: Said 'how many times do you to intercourse?' (not appropriate))*
>
> [Examiner] said not appropriate. I don't see this as not appropriate. [A5, S4]
>
> *(In response to: Didn't use the pictograms all the time (sign language, facial expressions, etc…. good but was easier for patient to understand pictograms)*
>
> Again, that could be interpreted as being a positive… I'm seeing the assessor 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 the tone was not warm? Does [the student] need to hug the doctor or what? [A5, S14]

Professional Perspective of Real-Life Practice

Assessors that assumed the professional perspective interpreted comments by relating meaning to what would occur in 'real life'. 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]

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

Narrative Comment	Response from 'Student Perspective'	Response from 'Professional Perspective'	Response from '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	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	I can see that the pharmacist is anxious. [A5]

	uncomfortable that she has to talk about this topic. [A3]	to trust you or be interested in communicating with you? [A10]	
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 exam 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].

Discussion

The purpose of this study was to explore how assessors give meaning to narrative data of student communication skills when interpreting an aggregated set of written comments obtained from a summative OSCE. 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 occurred during the student-patient interaction. In addition, assessors seemed to take different perspectives when interpreting assessment comments, i.e. the perspective of the student (places him or herself in the shoes of the student), the examiner (adopts the role of the examiner and re-interprets comments according to their own standards or beliefs), or the professional (acts as the profession's gatekeeper by considering the assessment to be a representation of real-life practice). Our findings show that these differences in assessor perspectives may affect assessors' explanation, framing or understanding of assessment comments. As such, our findings may contribute to further understanding of assessor variability in interpretation and use of narrative assessment data in judgement and decision-making of student performance.

Our findings suggest, in line with previous research, that an assessor's approach and perspective taken to interpretation of rich data is influenced by their beliefs, interpretations of the assessment setting, and personal performance theories.[11-13] For example, an assessor who assumes the perspective of the student when interpreting comments may do so because of their beliefs that the assessment task and 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 conceptualizations of what is to be considered effective task performance behaviour. Assessors who assume the perspective of the professional may do so because they feel it is their role to be the gatekeeper to the profession and because they believe the OSCE is an authentic representation of real-life situations, allowing for extrapolation of interpretations to professional practice. Based on these findings, it is conceivable that different assessor perspectives may result in variability in performance judgement and

decision-making.[14] The assessor that takes the student perspective may be more lenient (forgiving) in scoring or judgement, for instance, as compared to the assessor who takes the professional perspective and considers him or herself to be the profession's gatekeeper – feeling the need to be very strict. Further research, however, is needed to investigate to what extent assessors' perspectives influence how an assessor judges student performance in the context of a summative OSCE or other standardized assessment.

Our findings are also in line with previous research showing that assessor variance may not be the result of what assessors focus on but instead, may be due to the ways assessors interpret and bring meaning to data.[12,13,15] Our study adds to this argument by identifying and describing two different strategies (comparing/contrasting and making a mental image) and three different perspectives that assessors may adopt when conceptualizing student performance based on narrative assessment data in a high stakes assessment context. Previous research suggests that the assessment context affects the ways in which assessors approach narrative data.[13] As such, the assessment format (standardized or unstandardized) and the assessment goals (formative or summative) could be factors that influence the perspective an assessor brings to the data, and further research is needed to examine the role of context in narrative data interpretation. Although our data suggest that every individual assessor in our study largely interprets narrative assessment comments through adoption of a predominant perspective, further research is needed to investigate if and when, or under which conditions, assessors switch perspectives in interpretation of narrative assessment data.

Limitations

This study should be interpreted in light of some limitations. First, this was a single-centered study and so the types of comments, context for 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 previous participation in OSCE assessment. Though the study was limited to the State of Qatar, findings may be of relevance to others and justify future study in different settings. 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 by the assessors when comparing and contrasting or visualizing the data, we did this deliberately to reflect how assessors would likely receive a set of narrative comments in real-life practice and believe any effect would only negligibly impact our results regarding assessor idiosyncrasies in interpretation of narrative assessment data. Despite these limitations, we therefore feel our findings are important for understanding how assessors interpret/use narrative data and help to understand assessor variability in performance assessments.

Implications

This study builds on previous research relating to the interpretation of rich performance data and has implications for assessment practice. Our findings support arguments that the key to credible assessment is not in stripping or reducing data into smaller or simplified components, nor through standardization or assessor training, but instead we must develop

ways to cope with variance and ambiguity arising from differing interpretations of rich data.[15] Previous calls for the formation of clinical competency committees, where multiple assessors review rich data to make high-stakes decisions, may be a good start to account for the variation we (and others) have observed, and to allow for open and honest discussions about the different perspectives assessors bring to the data itself. Our results also support the need to pay attention to the factors related to assessor variance in assessor training and feedback. Despite the limited nature of what formal training can achieve, recognizing assessor beliefs/performance theories and accounting for these in coaching or feedback may promote reflection and awareness of the strategies/perspectives an assessor uses in judging performance. Our findings also support the current movement towards programmatic assessment, where multiple data points (such as narrative comments) are collected over multiple assessment contexts with the intent to provide both the learner and decision-makers with feedback about the learner's performance and progression.[1,2] Involvement of multiple assessors at all stages of interpretation and judgement of performance, with intentions to account for differing assessor perspectives, is necessary to not only ensure robustness in high stakes decisions, but may also add to the richness of assessment data for the purposes of understanding student performance across assessment settings.

Conclusions

This study explored how assessors interpret narrative data obtained from summative OSCEs. It was found that assessors bring different perspectives to narrative comments, which appear to influence 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 rich data. Results from our study can be used to enhance our understanding of the assessment process, in order to inform development and refinement of assessment procedures for the collection and interpretation of rich performance data. Based on our results, we can conclude that multiple assessors should interpret narrative data (when used for decision-making), in order to account for potential variation in grading arising from different assessors' perspectives.

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 App Res Mem Cogn. 2018;7:177-88.

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;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. Pat Educ Couns. 2012;88:106-12.

6. Harrison CJ, Molyneux AJ, Blackwell S, Wass VJ. How we give personalised audio feedback after summative OSCEs. Med Teach 2015;37(4):323-26.

7. Ginsburg S, Regehr G, Lingard L, Eva K. Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ. 2015;49:296-306.

8. Driessen E, van der Vleuten V, 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;39:214-20.

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;88:1539-44.

10. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr, G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. Acad Med. 2010;85:780-6.

11. Govaerts MJB, van de Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ. 2013;18:375-96.

12. Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. Adv in Health Sci Educ. 2018;23:275-87.

13. Berendonk C, Stalmeijer R, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. Adv in Health Sci Educ. 2013;18:559-71.

14. 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.

15. Schutz A, Moss PA. Reasonable decisions in portfolio assessment: evaluating complex evidence of teaching. Educ Policy Analysis Archives 2004;12:33.

16. Van Semeren MW, Barnard YF, Sandberg JAC. The think aloud method: A practical guide to modelling cognitive processes. Vol 2. London: Academic Press; 1994.

17. Sobh AH, Austin Z, Izham MIM, 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;9:1091-8.

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;5:1-12.

Chapter 5

Reproducibility of narrative assessment data on communication skills in a summative OSCE

Published as: Wilby KJ, Govaerts MJB, Dolmans DHJM, Austin Z, van der Vleuten C. Reproducibility of narrative assessment data on communication skills in a summative OSCE. Pat Educ Counsel 2019 (early online). DOI:10.1016/j.pec.2019.01.018

Abstract

Objective
To quantitatively estimate the reliability of narrative assessment data regarding student communication skills obtained from a summative OSCE and to compare reliability to that of communication scores obtained from direct observation.

Methods
Narrative comments and communication scores (scale 1-5) were obtained for 14 graduating pharmacy students across 6 summative OSCE stations with 2 assessors per station who directly observed student performance. Two assessors who had not observed the OSCE reviewed narratives and independently scored communication skills according to the same 5-point scale. Generalizability theory was used to estimate reliability. Correlation was used to evaluate the relationship between scores from each assessment method.

Results
A total of 168 narratives and communication scores were obtained. The G-coefficients were 0.571 for scores provided by assessors present during the OSCE and 0.612 for scores from assessors who provided scores based on narratives only. Correlation between the two sets of scores was 0.5.

Conclusion
Reliability of communication scores is not dependent on whether assessors directly observe student performance or assess written narratives, yet both conditions appear to measure communication skills somewhat differently.

Practice Implications
Narratives may be useful for summative decision-making and help overcome the current limitations of using solely quantitative scores.

Chapter 5

Introduction

Patient-centered communication is a core competency for health professionals and can be largely responsible for informing public perception of whether or not one is perceived to be a good practitioner.[1,2] Practitioner-patient communication is also known to directly impact patient health outcomes.[3-5] As such, health professional training programs must develop students to be good communicators. Summative assessment can aid programs in doing so by informing competency-based decisions of student communication competencies prior to entering practice.[6,7] However, any decision based on such assessment must arise from assessment methods that are both credible and defensible.

Objective Structured Clinical Examinations (OSCEs) are widely popular as a summative assessment method to test communication competencies within a standardized or simulated environment.[7] Competency-based decisions from OSCEs are typically based on numbers and scores derived from rating scales, checklists, or rubrics.[8-10] However, rubrics and checklists may not capture pertinent student behaviours that aid in the understanding of overall performance and may help to ensure credible and defendable assessment decisions.[10-12] These assessment tools typically identify competencies and sub-competencies that assessors must observe, synthesize, and judge yet studies have shown that assessors do not always conceptualize performance in line with the framework outlined by the tools themselves.[12,13] Furthermore, assessors may differ in their reasoning when making judgements based on what they observe, capturing of which may result in rich information that allows greater understanding of how a student's behaviours and actions are perceived and interpreted across different assessors and different contexts.[12] As a result, there are increasing calls to reform assessment tools, in order to better capture the quality of student task performance, provide more meaningful data for decision-making, and obtain rich feedback for student learning.[10,12] In order to meet these requirements, generation of qualitative data (i.e. narratives) is becoming increasingly important, as it provides a mechanism to capture and relay important contextual performance information to program directors or others who are ultimately in charge of making competency-based decisions.[10,14,15]

Although the use of narratives as an assessment tool is gaining credibility, questions remain regarding utility for summative decision-making. In particular, it is largely unknown how well assessment based on narrative data can discriminate between good and poor performance in different contexts. Previous research in workplace settings suggests narrative provides a strong enough 'signal' for assessors to reliably discriminate between levels of trainee performance when making judgements based on narrative alone and that judgements based on narrative demonstrate superior reliability, as compared to scores.[16,17] In these settings, supervisors work with residents over prolonged periods of time and are generally required to judge overall clinical competencies. Little is known, however, regarding the utility of narrative assessment methods for high-stakes standardized assessments, where interactions are typically one-time, short, and assessed by faculty who may have no prior knowledge of student capabilities. As such, the purpose of this study was to explore to what extent narrative obtained from these assessments provides enough 'signal' to reliably discriminate between levels of student performance.

The aims of this study were the following:

1. To estimate the reliability of narrative assessment data regarding student communication skills obtained from a summative OSCE
2. To compare the reliability of narrative assessment data to communication scores obtained from direct observation during the OSCE
3. To evaluate the relationship between the two scores in order to assess alternate form reliability between the different assessment methods

Methods

Study Design

We used a quantitative approach to exploring the reliability of narrative assessment data. Our study was designed in line with a previous study that used a similar methodology to estimate reliability of narrative comments from in-training evaluation reports.[16] Data were obtained as part of a summative OSCE for graduating pharmacy students. Assessors who wrote narrative comments and scored communication did so outside of normal grading practices and narrative data were only provided to students if requested. Generalizability theory was used to estimate reliability coefficients, as it allows for disentangling sources of error across multiple facets (student, station, assessor), as compared to other measures of reliability (e.g. inter-rater reliability) that do not.[18] In other words, it allows for separation of the signal (i.e. variance attributed to differences between candidates) from the noise (i.e. error resulting from other facets).[19] The study was exempted from full ethical review by the Qatar University Institutional Review Board (QU-IRB 883-E/18).

Setting / Context of the study

The study was conducted at the College of Pharmacy in Qatar University. The College of Pharmacy has an undergraduate Bachelor of Science in Pharmacy program and a post-graduate Doctor of Pharmacy program that are accredited by the Canadian Council for Accreditation of Pharmacy Programs (CCAPP).[20] As part of regular educational requirements, all graduating students from the Bachelor of Science in Pharmacy program are required to complete a summative OSCE that is blueprinted to the program's exit-from-degree competency framework (AFPC).[21] Each included station requires students to interact with a standardized patient or health professional to solve a patient's drug therapy problem. Students receive a brief description of each station upon entering the room and are provided with standardized hardcopy drug information references. Robust procedures for case development and validation were adapted from a Canadian model and described previously.[22,23] Our study included all stations from the 2016 OSCE that were designed to assess communication skills (n=6).

Participants

Fourteen students were recruited via email from a total population of 28 students. All students were taking part in the OSCE exam that was scheduled as part of their curriculum. Those recruited, however, agreed to have additional data collected and analyzed according to the study protocol. All students were female and completing the last month of study before graduation from the program. For writing of narratives, 12 assessors were recruited to evaluate communication skills during the 2-cycle OSCE (two assessors per station). These assessors were in addition to the assessors present during the exam to score students

according to normal procedures. Assessors were eligible to participate if they were a health professional, trained in assessment of communication, and if they had experience assessing student communication skills during previous OSCEs. These assessors were further trained during a 1-hour group session in advance of the OSCE by explaining study objectives and providing samples of narrative assessment comments extracted from the literature.[24] These examples were not related to communication, so as to prevent anchoring or seeding bias during the actual study.

After completion of the OSCE, two additional assessors were recruited to score communication skills solely based on narratives. These additional assessors were from the same assessor pool and had the same experience assessing communication skills as the other assessors recruited to score students and write narrative. These assessors were not involved in the OSCE, and did not directly observe student performance in communication. They were provided with a 30-minute introductory meeting to explain study objectives and procedures.

Research procedures and data collection

Step 1: assessment of communication skills based on direct observation of student performance during the OSCE. Assessors remained constant for each station throughout both OSCE-cycles and they were asked to write narrative comments of students' communication skills during the observed interaction. Evaluations were handwritten on a blank sheet of paper. Instructions to assessors were: "Please use the space below (and on the reverse if needed) to write a detailed narrative evaluation of the students' communication skills". No direction in the length or content of assessment comments was purposefully given, in order to minimize bias in terms of the skills, behaviours, and other attributes that assessors focus on. Assessors were given 17 minutes to write narrative assessment comments per student (8 minutes of observation, 9 minute break). Assessors were instructed to keep the narrative comments strictly confidential from their co-assessor or any other person to avoid data contamination. Assessors were also asked to assign an overall performance score to each student according to a 5-point communication assessment scale with anchors at 1, 3 and 5 points (1 = Communicates inappropriately and ineffectively to the task, 3 = Communicates with some logic and comprehension but not applied consistently, 5 = Communicates precisely, logically and perceptively to the encounter, integrating all relevant components). The global rating scale used for communication scoring in this study was previously validated and studies have shown good psychometric properties.[23,25] All assessors were familiar with the tool and had been previously trained using the tool via pre-assessment calibration exercises and post-assessment debriefing. No instruction was given to assessors regarding the order in which they completed the assessment tasks. Scores for each assessor pair per student were combined into one composite score for analysis.

Step 2: scoring of communication skills based on narrative assessment data. Upon completion of the OSCE, the two additional assessors were provided with the full narrative sets for each station. The communication scores as given by the OSCE assessors (Step 1) were not provided. Each assessor independently reviewed all individual narratives and assigned a communication score according to the generic assessment scale described above. These assessors were also instructed to not communicate with each other about the

narratives during the scoring procedure, which lasted approximately 3 hours. Once complete, the two scores obtained for each narrative were combined into a final composite (summed) score.

Data Analysis

The final data set consisted of composite (summed) communication scores obtained from the OSCE assessors (step 1), in addition to composite (summed) scores from the assessors who scored the narratives (step 2). Scores were stratified per station, entered into excel, and checked for errors. Means with standard deviations were used to summarize each set of scores. Correlation between composite scores obtained during the OSCE and composite scores based on narrative was determined using Spearman's rank correlation coefficient. For communication scores obtained during the OSCE, a G-study was conducted with students crossed with stations by assessors nested in stations [Px(R:S)]. The object of measurement was student communication scores. Facets included stations and assessors who scored communication during the OSCE. For communication scores based on narrative alone, the same study design was used [Px(R:S)], with 'R' representing assessors who scored communication based on narrative. For each of the G-studies, follow up decision studies were completed to determine the number of stations required to achieve G-coefficients of 0.80. All statistical analyses were completed using G_string.[26]

Results

General results

All 14 recruited students completed all six stations, resulting in 168 total communication scores with narratives. An example of a narrative is provided in Box 1.

Box 1. An example of narrative obtained from the OSCE

Narrative Examples
Example 1: The student used some terminologies that are quite strong for a listener/ patient (e.g severe …) her English language is a bit weak but still it could be understood. Her tone is quite loud and she sometimes lowers her voice and sometime "shout". She was not organized in her thoughts and just waits to listen to whatever the patient needs and answers accordingly. She felt uncomfortable and lost and kept looking at assessors. Regardless of the fact that she provided right recommendation, she had enough time to counsel the patient about what she missed in order to ensure safety. She showed some empathy towards the end but didn't maintain good eye contact or non verbal gestures with the patient. Example 2: • Very attentive to the patient with good eye contact • Very good voice projection • Courteous – shows empathy with the patients condition • Communication is well-structured and tailored to the patient's condition and questions • Very good variation of voice tone • She efficiently managed to adapt her communication to address

patient's concerns
- Overall, the student was polite and very pleasant
- Good body language
- Very good listener

The mean (standard deviation) of communication scores obtained during the OSCE was 3.64 (0.75) and 3.54 (0.86) from scores based on assessors reading narrative alone. The mean, standard deviation, and standard error of the mean per station are provided in Table 1.

Table 1. Descriptive results of scores obtained or each station

Station	1	2	3	4	5	6
Mean score during OSCE (SD)	3.3 (0.71)	3.7 (0.77)	3.7 (0.71)	3.9 (0.92)	3.6 (0.62)	3.6 (0.63)
SEM	0.13	0.15	0.13	0.17	0.12	0.12
Mean score from narrative (SD)	3.1 (0.76)	3.5 (0.93)	3.6 (0.76)	4.0 (0.89)	3.6 (0.88)	3.4 (0.74)
SEM	0.10	0.12	0.10	0.12	0.12	0.10

SD = standard deviation
SEM = standard error of the mean

Results for Aim 1: To estimate the reliability of narrative assessment data regarding student communication skills obtained from a summative OSCE

Table 2 provides variance components and G-coefficients for scores obtained based on narratives. The variance component for persons (P) accounts for about 14% of the total variation in scores. The narrative assessors nested in stations variance component (R:S) accounts for approximately 8% of total variance. As can be seen from table 2, the residual variance component contributes most to score variance (72%). The G-coefficient based on having two assessors score all narrative after completion of the OSCE was 0.612. The number of stations required to reach a G-coefficient of 0.80 was 15.

Results for Aim 2: To compare the reliability of narrative assessment data to communication scores obtained from direct observation during the OSCE

Table 2 also provides variance components and G-coefficients for communication scores obtained from assessors present during the OSCE. The variance component for persons (P) accounts for about 11% of the total variation in scores. The assessors nested in stations variance component (R:S), accounts for about 5% of total variance. As can be seen from table 2, the residual variance component contributes most to score variance (83%). The G-coefficient based on having two assessors per station and six stations was 0.571. The number of stations to reach a G-coefficient of 0.80 was 18.

Table 2. Variance components, G-coefficients, and results from D-studies

Variance Component (VC)	Scores Based on Direct Observation	Scores Based on Narratives
Effect VC (p)	11%	14%
Effect VC (s)	0.7%	6%
Effect VC (r:s)	5%	8%
Effect VC (residual)	83%	72%
G Coefficient	0.571	0.612
Number of stations to reach G Coefficient of 0.80	18	15

p = student communication ratings (object of differentiation)
s = station
r:s = rater nested in station
r = rater

Results for Aim 3: To evaluate the relationship between the two assessment methods
The correlation coefficient between composite scores obtained during the OSCE and composite scores based on narrative was 0.50 ($p<0.05$).

Discussion

In this study, we used a quantitative approach to investigate reliability of narrative assessment data obtained from a summative OSCE. Our findings suggest that reliability is similar when two assessors judge communication skills either by direct observation during an OSCE or by reviewing and interpreting narrative comments. Although generalizability coefficients >0.8 are typically recognized as 'excellent' markers of reliability, our estimate (0.612) supports the notion that the 'signal' in narratives enables assessors to conceptualize student performance and discriminate between good and poor performers. More specifically, it appears from our results that narrative offer similar discriminatory ability, as compared to scores based on direct observation of performance within the OSCE (generalizability coefficients of 0.612 and 0.571, respectively).

Findings from our study are comparable to findings from previous studies on reliability of communication scores for OSCEs.[27,28] Reliability coefficients obtained from communication scales in these studies were commonly lower than 0.8, which is generally considered too low for high stakes decision-making.[27,28] Reasons for this may be numerous but it is suggested that context, including OSCE content, likely plays a very central role when scoring communication competencies. Achieving high reliability is therefore dependent on testing students on a large sample of stations[15,27] and, as a consequence, is difficult to achieve due to reasons of feasibility and resource availability. Findings from our study suggest the reliability of narratives obtained from OSCEs is relatively low compared to reliability of narratives obtained in workplace settings. Ginsburg and colleagues, for example, reported reliability coefficients > 0.8 for narrative assessment data from in-training evaluation reports for eight or more reports.[16] However, these findings were based on four assessors.

Furthermore, the Ginsburg study focused on overall clinical competence, which may not only have influenced the type of comments and language used to convey judgements, it may also have inevitably resulted in assessors sampling performance information over longer periods of time and sampling across multiple competency domains, contributing to richness of data that could be taken into account when assessing trainee performance.[16]

The correlation between scores obtained from direct observation and those obtained from narratives was moderate at r=0.50. This moderate correlation suggests that while performance information captured by communication ratings and narratives is largely similar, assessment data also measure different aspects of student performance. Our findings may thus add to the evidence for the utility of narrative assessment data for summative purposes, as the combination of quantitative (scores) and qualitative (narrative) assessment data may result in more robust decision-making on the basis of rich information. However, it is difficult to interpret correlations between measures, as many factors could contribute to discrepancies observed, such as assessor characteristics, sample size, OSCE content, etc. For example, it could be a result of assessor tendencies to include constructive criticism and describe areas for improvement in comments, as opposed to scores. While it could be that narratives were measuring different aspects of performance, it could also be a result of differences in interpretation of performance by the assessors recruited to score narrative.[29] This finding, therefore, may warrant further study to better understand differences in assessment data between direct observation and narrative comments, as well as to explore the role of each in the context of overall decision-making.

If narratives are to be used for assessment purposes within OSCEs, a key point moving forward will be to investigate which aspects of student performance are lost or gained when using different assessment methods (e.g. interpretation of narratives vs. scores based on direct observation). We know from previous literature that assessors are influenced by many factors when observing performance and may have difficulty distinguishing between competencies (e.g. distinguishing between 'application of medical knowledge' from 'communication about health care issues in patient care'), which may impact communication skill assessment.[30,31] Asking assessors to limit their narrative to a single competency domain may therefore result in assessment data that are specific and meaningful indeed, yet do not entirely capture assessors' holistic, integrative judgement of the student-patient communication. Alternatively, global ratings may represent and include judgements on construct-irrelevant elements in task performance resulting from idiosyncrasies present in assessors' perceptions and interpretations. In order to gain a better understanding of these considerations for narrative assessment, future studies are required to investigate how narratives capture competency-based student performance data, how an assessor interprets the data to form an overall impression of student competence, and how much data an assessor needs (i.e. saturation) to inform a credible performance decision.

Limitations

The results of this study should be interpreted with consideration of some limitations. First, the sample size of students was small and the population was relatively homogenous. While this likely influenced the reliability coefficients obtained, it should be noted that the number of narratives used in the data set was actually quite large (n=168). In fact, each evaluation

contained on average 9 phrases representing a different idea or opinion, resulting in over 1,500 phrases to read and interpret. Secondly, the procedures employed provided assessors with 17 minutes to write narratives, as it was anticipated that it would take longer than the actual station time of 8 minutes to write comments. Not only did this reduce the sample size, but it also raises concerns regarding the practicality of writing narrative during OSCEs. However, it should be noted that assessors felt 8 minutes was sufficient time to evaluate students and write comments. Thirdly, our results showed a large amount of general (residual) error with a smaller proportion coming from students and the other facets. Despite being a limitation, similar results have been found in other studies.[19,32]

Conclusion

The results of our study further our understanding of the utility of narrative within assessment procedures during OSCEs. Scoring of narratives resulted in similar reliability of student communication performance scores. Reliability does not seem to be dependent on whether assessors directly observe the student-patient interaction or assess written narratives. However, scores from each of these conditions appear to measure communication skills somewhat differently. As such, further investigation into the utility of narratives for assessment of communication skills during OSCEs is warranted.

Practice Implications

This study demonstrated similar moderate reliability of communication scores obtained from direct observation during an OSCE with scores obtained based on narrative comments of student performance. This finding shows that assessors are able to read narrative comments and assign scores in a relatively discriminatory manner, similar to that of watching student performance live. As such, narrative evaluations of student communication skills obtained during OSCEs may support summative assessment practices by providing a rich data source with discriminatory power for competency decision-making. This finding has implications for programmatic assessment, which calls for rich sources of data across multiple assessment contexts as a student moves through a training program.[33] The reliability demonstrated for narratives in our study shows that narrative obtained from summative OSCEs may provide program administrators or clinical competency committees with reliable and rich data, as compared to scores alone, that can inform judgements and support decision-making. For example, narrative comments and scores from OSCEs could be assessed together with different data points obtained across other assessment contexts (workplace-based training evaluations, reflective assignments, practical laboratory assessments, etc.) to inform decisions for pass-fail, promotion, or need for remediation. Before implementation in practice, however, research must inform how individuals or committees interpret aggregated narrative data (e.g. across all stations) and how inclusion of narrative data in programmatic assessment may influence judgements. Furthermore, our study findings also suggest that narrative assessment and direct observation may provide different insights into student communication skills. Until we have better understanding of the similarities and differences between these two assessment methods, the use of both methods should be encouraged to a) ensure robust decision-making and b) provide meaningful data for remediation (performance development).

References

1. Papp R, Borbas I, Dobos E, Bredehorst M, Jaruseviciene L, Vehko T, Balogh S. Perceptions of quality in primary health care: perspectives of patients and professionals based on focus group discussions. BMC Fam Pract 2014;15:128.

2. Doubova SV, Guanais FC, Perez-Cuevas R, Canning D, Macinko J, Reich MR. Attributes of patient-centered primary care associated with the public perception of good healthcare quality in Brazil, Colombia, Mexico, and El Salvador. Health Policy Plan 2016;31:834-43.

3. Stewart MA. Effective physician-patient communication and health outcomes: a review. Can Med Assoc J 1995;152:1423-33.

4. Mercer SW, Higgins M, Bikker AM, et al. General practitioners' empathy and health outcomes: a prospective observational study of consultations in areas of high and low deprivation. Ann Fam Med 2016;14:117-24.

5. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky K, Riess H. The Influence of the Patient-Clinician Relationship on Healthcare Outcomes: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. PLoS One 2014;9:e94207.

6. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32:676-82.

7. Epstein RM. Assessment in medical education. N Eng J Med 2007;356:387-96.

8. Setyonugroho W, Kennedy KM, Kropmans TJ. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. Pat Educ Couns 2015;98:1482-91.

9. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalizability. Med Educ 1996;30:38-43.

10. 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. Pat Educ Couns 2012;88:106-12.

11. Driessen E, van der Vleuten V, 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;39:214-20.

12. 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.

13. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr, G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. Acad Med 2010;85:780-6.

14. Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? Med Educ 2007;41:1121-23.

15. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.

16. 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;92:1617-21.

17. Bartels J, Mooney CJ, Stone RT. Numerical versus narrative: A comparison between methods to measure medical student performance during clinical clerkships. Med Teach 2017;39:1154-8.

18. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006-12.

19. Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE guide no. 68. Med Teach 2012;34:960-92.

20. Canadian Council for Accreditation of Pharmacy Programs, Accredited Programs. http://www.ccapp-accredit.ca, 2017 (accessed 22 July 2018).

21. Association of Faculties of Pharmacy of Canada, Educational Outcomes for first professional degree programs in pharmacy in Canada – June 4, 2017. http://afpc.info/node/39, 2017 (accessed 22 July 2018).

22. Wilby KJ, Black EK, Austin Z, Mukhalalati B, Aboulsoud S, Khalifa SI. Objective structured clinical examination for pharmacy students in Qatar: cultural and contextual barriers to assessment. East Med Health J 2016;22:251-7.

23. Sobh AH, Austin Z, MI Izham M, 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;9:1091-8.

24. Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med 2012;87:419-29.

25. 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;5:1-12.

26. McMaster Education Research, Innovation & Theory, G_String. http://fhsperd.mcmaster.ca/g_string/, 2015 (accessed 22 July 2018).

27. Brannick MT, Tugba Erol-Korkmaz H, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011;45:1181-9.

28. Comert M, Zill JM, Christalle E, Dirmaier J, Hartner M, Scholl I. Assessing communication skills of medical students in Objective Structured Clinical Examinations (OSCE) – A systematic review of rating scales. PLoS ONE 2016;11:e0152717.

29. Ginsburg S, Regehr G, Lingard L, Eva K. Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ 2015;49:296-306.

30. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32:676-82.

31. Eva KW. Cognitive influence on complex performance assessment: Lessons from the interplay between medicine and psychology. J App Res Mem Cogn 2018;7:177-88.

32. Govaerts MJ, van der Vleuten CP, Schuwirth LW. Optimising the reproducibility of a performance-based assessment test in midwifery education. Adv Health Sci Educ Theory Pract 2002;7:133-45.

33. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach 2011;33:478-85.

Chapter 6

Discussion

Discussion

The need to refine our performance-based assessment approaches in health professionals' education is clear. With the rise of competency-based education, assessments need to generate data on student performance that support the dual purposes of assessment for and of learning.[1,2] Traditional assessment approaches of reducing student performance to numbers or standardized descriptors do not allow for capturing and understanding variations and idiosyncrasies in assessors' interpretations of task performance.[3] Shifting our focus to obtain rich data (e.g. narrative) from assessments, including OSCEs, may overcome these limitations and support credibility of performance-based judgements and decision-making. However, in order to ascertain usefulness of narrative assessment data for fair (i.e. credible and defensible) decision-making, we must further our understanding of the role of the assessor in narrative assessment systems by exploring how they process information about student performance. More specifically, we need to explore what assessors actually observe and select as relevant performance data, and how they interpret performance information during direct observation of task performance or when reviewing rich data obtained from the assessment. Given the important role of OSCEs in competency-based assessment programmes, and in order to better understand assessor judgements in these standardized assessment settings, this PhD thesis therefore aimed to explore the following research questions:

1. How do assessors process performance data when judging student communication performance in OSCEs?
    a. What do assessors pay attention to when judging student communication in OSCE stations and how is this influenced by assessor characteristics?
    b. How do assessors convey their observations and interpretations into narrative assessment data?

2. How does use of narrative assessment data impact assessment quality?
    a. How do assessors interpret narrative assessment data (provided by others)?
    b. How reliable are scores based on narrative data obtained from an OSCE, as compared to scores based on direct observation?

In the following sections, we will first answer our research questions based on the data contained within this thesis and then discuss practical implications, directions for future research, and the strengths and limitations of this thesis.

How do assessors process performance data when judging student communication performance in OSCEs?

In order to answer this research question, we focused on two key steps in assessors' information processing: how assessor characteristics (i.e. cultural orientations) affect what assessors pay attention to and value when observing task performance and how OSCE-assessors formulate their observations and interpretations into rich data (e.g. narratives).

We began in Chapter 2 by investigating how assessors' cultural orientations are related with interpretation of student performance as portrayed in three videotaped OSCE scenarios. Our key finding from this study was that assessors from various cultural backgrounds generally agreed on what constitutes good or poor communication performance, whereas the influence of cultural orientations on communication scores and values of communication behaviours seemed to be scenario dependent and more prevalent within borderline performance. For example, those assessors scoring high on masculinity paid more attention to positive aspects of affective communication and provided a higher (more favourable) overall global communication score when observing the borderline performance video. This finding aligns with previous research showing that practitioner-patient communication itself is influenced by cultural orientations but further demonstrates that these cultural communication preferences also translate into assessor preferences for interpreting and judging communication performance in (standardized) assessment settings.[4] Our findings thus add to existing data that assessor characteristics may be a key factor for assessors' interpretation and scoring of student performance.[5] Assessor characteristics such as assessor expertise or clinical competence, but especially assessors' cultural orientations, may be an important factor influencing interpretation and judgement of communication when observing borderline communication performance. From that perspective, these findings are in line with and contribute to assessor cognition research suggesting that assessors may very well hold different perspectives that represent real-world practice, and bring varying yet equally valid interpretations of candidate performance.[6]

By demonstrating the presence of differing (i.e. idiosyncratic) assessor judgements in OSCEs, the results from Chapter 2 thus called for further studies to make assessors' reasoning in OSCEs explicit and to investigate how assessors formulate what they observe into narrative assessment comments. In Chapter 3, we therefore recruited assessors to write narrative assessment comments about student communication skills for a summative OSCE. We sought to explore the skills/behaviours assessors focus on when writing comments and the linguistic strategies they may use to convey their observations and interpretations. Our primary finding was that assessors used four overarching constructs (i.e. confidence, adaptability, patient safety, and professionalism) to convey their interpretations of good/poor performance, rather than focusing on specific communication behaviours or skills (e.g. eye contact, facial expressions, etc.). This finding aligned with previous studies showing that assessors, and expert-assessors in particular, may focus more on the 'big picture', rather than deconstructing competence into distinct components.[7,8] We also found that assessors wrote their comments using elements of politeness (i.e. hedging) albeit less frequently than comments extracted from workplace based assessment.[9] The findings from this study show that the way in which assessors formulate comments and the constructs they use to convey observations and judgements are important for interpretation of what they perceive to be key elements of good or poor performance in communication. Our findings thus provide an initial framework of key constructs that can inform the development of better anchored assessment tools, which are constructively aligned with assessors' map of the reality of professional practice, potentially contributing to quality enhancement of communication assessment within OSCEs.

How does use of narrative assessment data impact assessment quality?

Equipped with greater knowledge of how assessors formulate what they observe into rich performance data from OSCEs, we decided to further our understanding of the quality of narrative assessment data by examining how assessors interpret narrative assessment data obtained in OSCEs and how use of narratives may impact assessment reliability. In Chapter 4, we investigated how assessors bring meaning to aggregated sets of narrative assessment comments written by others. Our findings suggest that assessors used similar strategies (i.e. comparing and contrasting, forming mental images) to analyze, integrate, and interpret data, but approach and interpret narrative comments from three distinct perspectives: the student (places him or herself in the shoes of the student), the examiner (adopts the role of the examiner and re-interprets comments according to their own standards or beliefs), or the professional (acts as the profession's gatekeeper by considering the assessment to be a representation of real-life practice). Each perspective had implications for how the assessor interpreted the comments and our findings demonstrated that the assessor is once again fundamental to how student performance data is interpreted. Assessors that assumed the student perspective, for example, may be more lenient in judgement/grading while assessors who assumed the professional perspective may be stricter. Findings from this study align with previous studies that show that interpreting rich performance information is largely dependent on how the assessor approaches the data and selects which data to process and evaluate.[10,11] Our study adds to this literature by proposing three specific perspectives that may be important for interpreting rich data but also shows that assessment quality (i.e. variance in assessor judgements) may be influenced by how assessors use these perspectives to interpret, judge, and eventually make decisions based on rich performance data.

Our results from Chapter 4 showed that assessors might differ in how they approach and interpret narrative assessment comments from OSCEs. Obviously, this may raise questions about the impact of assessor variance on reliability (reproducibility) of narrative assessment data, which is generally considered to be a prerequisite for use of assessment data for high-stakes decision-making. Our final study, presented in Chapter 5, aimed to explore this by comparing the reliability of scoring based on narrative data with scores obtained from direct observation of performance. Using generalizability theory, we found reliability coefficients between the two conditions to be similar (0.571 vs. 0.612 for direct observation and narrative, respectively) but the scores based on narrative data were only moderately correlated with the scores based on direct observation (r=0.5). Reliability of communication scores was therefore not dependent on whether or not assessors directly observed student performance or assessed written narratives, yet both conditions appeared to measure communication skills somewhat differently. Consequently, large numbers of OSCE stations and multiple assessors may be required to ensure assessment quality and reliability suitable for high stakes decision-making when using either assessment approach.

Implications

This thesis has both theoretical and practical implications. From a theoretical perspective, our findings build on previous assessor cognition research by providing greater insight into assessor variance and by explaining or providing rationale for why variance occurs during

communication assessment in OSCEs. Our studies contribute to existing literature that supports assessors as being meaningfully idiosyncratic, rather than attributing variance between assessors to be solely due to error.[6,12] Inherent characteristics, such as assessors' cultural orientations appear to drive and explain differences in how assessors interpret and judge communication. Our studies also provide additional explanation by showing that aside from assessor characteristics, such as cultural orientations, the way in which assessors approach performance data is important for how they interpret and judge performance. For example, different perspectives may very well affect an assessor's leniency (or strictness) in grading and decision making in performance assessments. Our findings thus point at the complexity of human (expert) judgements in assessment tasks, in that assessor characteristics may indeed result in meaningfully different, yet equally valid performance interpretations, as between-assessor differences may actually represent the way performance can and will be interpreted in real-life professional practice. From that perspective, between-assessor differences may provide meaningful information for student learning, as well as contribute to robustness of decision-making. Capturing between-assessor variance by having assessors write narrative comments, may provide a more authentic assessment that better reflects how a student will eventually be perceived as a practicing professional. Our findings, however, also show that assessor characteristics related to perspectives on performance assessments and performance data, may contribute to what is usually considered unwanted assessor bias or error (e.g. leniency). It is important to note that these between-assessor differences occurred despite employing well-trained, expert assessors in each of these studies.

If narrative is to be incorporated into assessment of communication within OSCEs, our findings have practical implications for assessment instruments and assessor training. First of all, as shown in Chapter 3, assessors do not always conceptualize performance as reflected in checklists or rubrics. This may call for refinement of these tools, in order to better reflect the key performance domains or constructs that assessors seem to focus on and value when judging performance. Recent evidence shows that aligning assessment with the assessor, instead of attempting to change an assessors' beliefs or focus, may better support the overall success and quality of performance assessment.[1] Secondly, in order to ensure credible decision-making based on performance data from OSCEs, instruments need to facilitate assessors' substantiation of judgments with a goal of obtaining a better understanding of how assessors conceptualize student performance. Refining assessment instruments to incorporate assessors' narrative data will capture assessors' differing perspectives and facilitate better understanding of any variance in scores or judgements.

Our results also have implications for assessor training. Despite using expert assessors with experience assessing student communication skills in OSCEs, we still found variability in terms of what they value when watching performance and the perspectives they bring to interpreting performance data. We therefore may need to rethink the way we train or coach assessors in future assessment design. Based on the findings from this thesis, the focus of training should likely shift towards making assessors aware of their own perspectives and beliefs and how these might impact their judgements and decision making, rather than focusing solely on training of assessment skills. If narrative assessment is to be adopted within OSCEs, training should also focus on assisting assessors to write meaningful data within narratives. Comments have to be meaningful for learning as well as decision

making. Capturing meaningful between-assessor variations in interpretations of performance will be useful for substantiating assessor judgements and better understanding student performance as a whole. Writing meaningful narratives, however, is cognitively demanding, and assessors need to develop the language and the skills to convey their own perspectives, interpretations and key messages within short periods of time.

Strengths and limitations

Some characteristics of our approach and findings highlight certain strengths of this thesis. First, we used a variety of methods to answer our research questions, including quantitative and qualitative approaches. We included methods, such as the use of stimulated recall, think-aloud and G-theory, targeted to investigate different aspects of assessors' information processing and the quality of narrative data. Second, our investigator team consisted of experts from a variety of disciplines (pharmacy, medicine, education), which allowed us to interpret data and make conclusions from broader perspectives. Thirdly, our findings add to the broader field of assessor cognition research by demonstrating that assessors' cultural orientations and use of different perspectives to approach performance data may inform how they interpret and ultimately judge student performance. As discussed above, these findings may have implications for assessment tool development and grading.

Some limitations of this thesis must be recognized. First, the sample sizes of student data obtained for our studies were relatively small and the population of students was fairly homogenous (gender, language, region of origin). Although this may have influenced the transferability of results, we were required to work within the program constraints. Secondly, all studies were solely focused on communication skills. It is therefore unknown if the same findings hold true for narrative written based on overall clinical performance or for other competencies being assessed during the OSCE. Finally, this thesis could only explore assessor processing from a few perspectives (observation and interpretation) and the influence of assessor processing on pass-fail decision-making in this context remains yet to be studied.

Suggestions for future research

Our findings justify further exploration of the role of the assessor in OSCEs and in particular, support conduction of studies aimed to evaluate how assessors make judgements and performance decisions based on narrative data obtained from OSCEs. In the studies included in this thesis, we explored assessors' values according to what they pay attention to, how they formulate their observations and judgements into words, and how they interpret narrative data written by others. Despite our findings suggesting that these processes may influence how an assessor forms judgements and performance decisions, we did not explicitly study the impact of between-assessor variance on pass-fail decisions. An obvious next step would be to determine how the differing perspectives an assessor takes when interpreting narrative data influences how they judge or grade performance. Another area for future research relates to our second research question and the concept of data

saturation. Although our findings showed a large number of stations would be needed to obtain high reliability when using narrative data to assess student performance, this result was based on judgements at the station level and based on conversion of narratives into scores. This, however, seems like an illogical step to take in narrative assessment systems. Future studies should therefore aim at investigating assessment quality using techniques and strategies from qualitative research, for instance to explore the number of narratives needed to achieve the 'data saturation' that is needed for assessors to arrive at a coherent interpretation of student performance.

Conclusion

The research presented in this thesis contributes to our understanding of the use of narrative data in OSCEs. Assessors appear to be fundamental to the assessment process and may potentially influence performance outcomes in terms of what they see when watching performance, how they formulate narrative comments based on performance, and how they interpret meaning of narrative comments written by others. These findings stress the importance of using narrative assessment data within OSCEs and support the notion that we must develop assessment instruments that account for assessor variance and use this data to better support assessment decisions.

References

1. Eva KW. Cognitive influence on complex performance assessment: Lessons from the interplay between medicine and psychology. J App Res Mem Cogn. 2018;7:177-88.

2. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356:387-96.

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. Meeuwesen L, van den Brink-Muinen A, Hofstede G. 2009. Can dimensions of national culture predict cross-national differences in medical communication? Pat Educ Cousel 75:58-66.

5. Berendonk C, Stalmeijer R, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. Adv in Health Sci Edu.c 2013;18:559-71.

6. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the black box differently: assessor cognition from three research perspectives. Med Educ. 2014;48:1055-68.

7. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ 2016;50:511-22.

8. Wilbur K, Hassaballa N, Mahmoud OS, Black EK. Describing student performance: a comparison among clinical preceptors across cultural contexts. Med Educ. 2017;51:411-22.

9. 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 in Health Sci Educ. 2016;21:175-88.

10. Oudkerk Pool A, Govaerts MJB, Jaarsma DADC, Driessen EW. From aggregation to interpretation: how assessors judge complex data in a competency-based portfolio. Adv in Health Sci Educ. 2018;23:275-87.

11. Schutz A, Moss PA. Reasonable decisions in portfolio assessment: evaluating complex evidence of teaching. Educ Policy Analysis Archives 2004;12:33.

12. Govaerts MJB, van de Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ. 2013;18:375-96

Summary

Summary

Chapter 1: Introduction

This PhD thesis began by examining the need to capture rich performance data that can substantiate assessor judgements and provide greater feedback for learning in performance-based assessments, including OSCEs. To date, OSCEs are primarily designed to be objective measures of student performance and typically use solely quantitative approaches to assessment. We reviewed what is known to date in this regard but also outlined calls for the collection of rich assessment data (e.g. narratives) within OSCEs. We furthermore explained how assessor cognition research has shown that assessors' performance judgements are inherently idiosyncratic and proposed that capturing judgement using narratives may support better understanding of performance and improve credibility and defensibility of assessment. Before rich assessment data can be used in OSCEs, however, we must better understand the role of the assessor in processing performance information from OSCEs and how assessment quality may be impacted by using narrative data. This led to the development of two central research questions:

1. How do assessors process performance data when judging student communication performance in OSCEs?
    a. What do assessors pay attention to when judging student communication in OSCE stations and how is this influenced by assessor characteristics?
    b. How do assessors convey their observations and interpretations into narrative assessment data?

2. How does use of narrative assessment data impact assessment quality?
    a. How do assessors interpret narrative assessment data (provided by others)?
    b. How reliable are scores based on narrative data obtained from an OSCE, as compared to scores based on direct observation?

Chapter 2

The study described in Chapter 2 focuses on determining what assessors pay attention to when observing student performance in OSCEs and how this may be influenced by cultural orientations. Twenty-five pharmacist-assessors watched 3 videotaped scenarios (patient-pharmacist interactions) and ranked each on a 5-point global rating scale. Videotaped scenarios demonstrated combinations of well-portrayed and borderline examples of instrumental and affective communication behaviours. We used stimulated recall and verbal protocol analysis to investigate assessors' interpretations and evaluations of communication behaviours. Uttered assessments of communication behaviours were coded as instrumental (task-oriented) or affective (socioemotional) and either positive or negative. Cultural orientations (power-distance, masculinity- femininity, uncertainty avoidance, and individualism-collectivism) were measured using the Individual Cultural Values Scale. Correlations between cultural orientations and global scores, and frequencies of positive, negative, and total utterances of instrumental and affective behaviours were determined. We found that both communication scores and utterances of communication behaviours were associated with cultural orientations (masculinity-femininity, power-distance) for a

video portraying borderline performance. Our findings thus confirm cultural orientation may be a source of assessor idiosyncrasy in interpretation of communication behaviours but may be dependent on the nature of the scenario portrayed. As such, our results for cultural orientations may reflect the broader importance of assessors' characteristics for explaining variation in how assessors process performance information and eventually decide on assessment scores/outcomes.

Chapter 3

Having determined that assessors' characteristics may influence what they pay attention to when observing student performance in OSCEs, in Chapter 3 we aimed to determine how assessors use narrative comments to convey impressions of good and poor performance. Eighteen assessors from Qatar University were recruited to write narrative assessment comments of communication skills for 14 students completing a summative OSCE. Assessors scored overall communication performance on a 5-point scale. Narrative evaluations for the top and bottom two performing students for each station (based on communication scores) were analyzed for linguistic strategies (deductively, according to politeness theory) and constructs (inductively, using grounded theory) that informed assessment decisions. We found the overarching constructs of confidence, adaptability, patient safety, and professionalism were key dimensions that assessors wrote to discriminate between student performance. Furthermore, we found that most comments were not written using an element of politeness but hedging was present in 22% of comments, and more commonly for poor performers. These results demonstrate that the assessor is fundamental to what data is selected to convey performance judgements using narrative and that this data may differ from traditional scoring tools or rubrics.

Chapter 4

After determining what assessors see (Chapter 2) and what assessors write (Chapter 3) with respect to student performance in OSCEs, in Chapter 4 we sought to explore how assessors interpret aggregated sets of narrative comments obtained from OSCEs. Narrative assessment comments of student communication skills as well as communication scores were obtained for 24 students across 6 stations of a summative OSCE. Aggregated narrative data across all stations was sampled for 9 students (3 good, 3 average, and 3 poor performers, based on communication score). For each of the students, ten expert-assessors reviewed the aggregated set of narrative comments. Cognitive (information) processing was captured through think-aloud procedures and verbal protocol analysis. We found expert-assessors primarily made use of two strategies to interpret the narratives, namely comparing and contrasting, and forming mental images of student performance. Assessors, however, appeared to use three different perspectives when interpreting narrative assessment comments: the student (places him or herself in the shoes of the student), the examiner (adopts the role of the examiner and re-interprets comments according to their own standards or beliefs), or the professional (acts as the profession's gatekeeper by considering the assessment to be a representation of real-life practice). These findings highlight the presence of assessor variance in interpretation of narrative assessment data, which may impact how assessors judge performance and therefore influence markers of assessment quality (e.g. reliability).

Chapter 5

In Chapter 5, we explored the influence of narrative assessment data on assessment quality by obtaining narrative assessment data from an OSCE and comparing reliability of communication scoring based on narrative versus direct performance observation. Narrative comments and communication scores (scale 1–5) were obtained for 14 graduating pharmacy students across 6 summative OSCE stations with 2 assessors per station who directly observed student performance. Two assessors who had not observed the OSCE reviewed narratives and independently scored communication skills according to the same 5-point scale. Generalizability theory was used to estimate reliability. Correlation was used to evaluate the relationship between scores from each assessment method. We found that reproducibility was similar for both conditions. The G-coefficients were 0.571 for scores provided by assessors present during the OSCE and 0.612 for scores from assessors who provided scores based on narratives only. These findings suggest reliability of communication scoring is not dependent on direct observation during the OSCE. Correlation between the two sets of scores was 0.5; suggesting narratives may have a role in measuring components of performance different to that of scores. Overall, these findings suggest narrative assessment comments may have value within OSCEs and support further exploration of the utility of narrative assessment comments in practice.

Chapter 6: Discussion

In Chapter 6, we answer our research questions by synthesizing the results of our studies and interpret meaning in relation to the existing literature. We also provide guidance for practical implications, suggestions for future research, and a short discussion of the strengths and limitations of this thesis. As an answer to our first research question: “How do assessors process performance data when judging student communication performance in OSCEs?” we can conclude that assessors’ information processing (i.e. what they pay attention to, how they translate observations into narrative data, and how they interpret narrative performance data) may influence how they judge and/or score performance. Our findings show that these processes may be influenced by a number of factors, including their cultural orientations and the perspective that they use to approach interpretation of performance data. In answer to our second research question: “How does use of narrative assessment data impact assessment quality?” we can conclude that despite differences in how assessors interpret narrative data, reliability of scoring is not dependent on directly observing student performance. Although, it appears that narrative data may measure communication in OSCEs somewhat differently than direct observation alone. These findings support further exploration of the use of narrative data in OSCEs, in order to substantiate assessor judgements and improve credibility and defensibility in scoring/judgement.

Samenvatting

Samenvatting

Hoofdstuk 1: Introductie

In het eerste hoofdstuk hebben wij de behoefte in kaart gebracht om bij vaardigheids- of competentietoetsen, waaronder stationstoetsen, rijke gegevens over de prestaties van studenten vast te leggen. Deze rijke prestatiegegevens kunnen oordelen van beoordelaars onderbouwen en betere feedback verschaffen aan studenten, ter bevordering van het leren. Tot op heden worden stationstoetsen voornamelijk ingezet voor een 'objectieve' meting van het vaardigheids- of competentieniveau van de student, waarbij doorgaans uitsluitend gebruik gemaakt van kwantitatieve benaderingen van beoordelen. Het hoofdstuk geeft een overzicht van wat er tot op heden over dit onderwerp bekend is, en beschrijft tevens de aanleiding en noodzaak om rijke beoordelingsgegevens (bijv. narratief, schriftelijk commentaar) te verzamelen bij stationstoetsen. Vervolgens wordt uitgelegd hoe onderzoek op het gebied van beoordelaarscognitie heeft aangetoond dat beoordelaars' prestatiebeoordelingen van nature persoonsgebonden (idiosyncratisch) zijn. Wij stellen daarom dat het opnemen van narratief commentaar in beoordelingen meer inzicht kan bieden in de prestaties van studenten en daarmee de geloofwaardigheid en verdedigbaarheid van de prestatiebeoordeling kan verbeteren. Voordat we echter gebruik kunnen maken van rijke beoordelingsgegevens bij stationstoetsen, moeten we beter begrijpen welke rol de beoordelaar speelt bij de verwerking van prestatie-informatie uit stationstoetsen en hoe het gebruik van narratieve gegevens de kwaliteit van de toetsing kan beïnvloeden. Op basis hiervan hebben wij twee centrale onderzoeksvragen geformuleerd:

1. Hoe verwerken beoordelaars prestatie-informatie bij het beoordelen van studenten op hun communicatieve vaardigheden tijdens de stationstoets?
    a. Waar letten beoordelaars op als zij studenten beoordelen op hun communicatieve vaardigheden tijdens de diverse stations van de toets en hoe wordt dit beïnvloed door beoordelaarskenmerken?
    b. Hoe brengen beoordelaars hun observaties en interpretaties tot uitdrukking in narratieve beoordelingsgegevens?

2. Hoe beïnvloedt het gebruik van narratieve beoordelingsgegevens de beoordelingskwaliteit?
    a. Hoe interpreteren beoordelaars (door anderen verschafte) narratieve beoordelingsgegevens?
    b. Hoe betrouwbaar zijn de scores die berusten op narratieve beoordelingsgegevens verkregen uit een stationstoets ten opzichte van scores die gebaseerd zijn op directe observatie?

Hoofdstuk 2

In de in Hoofdstuk 2 beschreven studie wordt aandacht besteed aan de vraag waar beoordelaars op letten bij het observeren van studentprestaties tijdens een stationstoets, en óf en hoe dit beïnvloed wordt door hun culturele oriëntatie. Vijfentwintig apotheker-beoordelaars bekeken drie op video opgenomen scenario's (van interacties tussen patiënten en student-apothekers) en waardeerden deze elk op een 5-punts globale beoordelingsschaal. De op video opgenomen scenario's toonden combinaties van instrumenteel en affectief communicatief gedrag, waarbij de effectiviteit van de diverse gedragingen per video varieerde (goed, slecht of twijfelachtig). We maakten gebruik van "*stimulated recall*" en verbale protocolanalyse om beoordelaars' interpretaties en evaluaties van communicatief gedrag te onderzoeken. Hardop uitgesproken interpretaties en evaluaties van communicatief gedrag werden gecodeerd als instrumenteel (taakgericht) of affectief (sociaal-emotioneel) en als positief of negatief. Culturele oriëntaties (machtsafstand, masculiniteit-femininiteit, onzekerheidsvermijding en individualisme-collectivisme) werden gemeten met behulp van de *Individual Cultural Values Scale* (schaal voor het meten van culturele waarden op individueel niveau). Correlaties tussen culturele oriëntaties en globale communicatiescores, alsook het aantal positieve, negatieve en het totaal aan door beoordelaars verwoorde evaluaties van instrumenteel en affectief gedrag werden berekend. We ontdekten dat zowel de communicatiescores als de verbale evaluaties van communicatief gedrag verband hielden met culturele oriëntaties (masculiniteit-femininiteit, machtsafstand) bij videofilmpjes waarin de student suboptimaal (twijfelachtig) communicatief gedrag vertoonde. Onze bevindingen bevestigen dus dat culturele oriëntatie mogelijk een bron is van beoordelaarseffecten (beoordelaarsspecificiteit) bij het interpreteren van communicatief gedrag, maar dat dit afhankelijk kan zijn van het soort scenario dat wordt weergegeven. In die hoedanigheid geven onze resultaten ten aanzien van culturele oriëntaties mogelijk een breder perspectief op het belang van beoordelaarskenmerken bij het verklaren van verschillen in de manier waarop beoordelaars prestatie-informatie verwerken en uiteindelijk tot communicatiescores en beslissingen over studentprestaties komen.

Hoofdstuk 3

Na te hebben vastgesteld dat beoordelaarskenmerken van invloed kunnen zijn op hetgeen waar zij op letten bij het observeren van studentprestaties tijdens de stationstoets, beoogden we in Hoofdstuk 3 te onderzoeken hoe beoordelaars narratief commentaar gebruiken om hun indrukken van goede en slechte prestaties over te brengen. Achttien beoordelaars van de universiteit van Qatar werd gevraagd om narratief beoordelingscommentaar te schrijven over de communicatieve vaardigheden van 14 studenten die een summatieve stationstoets doorliepen. Beoordelaars gaven daarnaast op een 5-puntsschaal een algemeen, globaal oordeel over de communicatieve vaardigheden van de student. Voor elk station werden de narratieve evaluaties van de twee best en de twee slechtst presterende studenten (gebaseerd op hun communicatiescores) geanalyseerd op gebruik van taalstrategieën (deductief, volgens de beleefdheidstheorie) en constructen (inductief, met behulp van gefundeerde theorie = grounded theory) die aan de beoordelingen ten grondslag lagen. We ontdekten dat de overkoepelende constructen "zelfvertrouwen", "aanpassingsvermogen", "patiëntveiligheid" en "professioneel handelen" de belangrijkste aspecten waren die beoordelaars hanteerden en opschreven om tussen studentprestaties te discrimineren. Voorts constateerden we dat het meeste commentaar

was geschreven zonder enige beleefdheidscomponent, maar dat er in 22% van het commentaar wel sprake was van vaag taalgebruik, met name in gevallen van slechte presteerders. Deze resultaten tonen aan dat de beoordelaar aan de basis ligt van welke gegevens er geselecteerd worden om prestatieoordelen in narratief commentaar vast te leggen, en dat dit kan verschillen van hetgeen in traditionele scoringsinstrumenten of *scoring rubrics* is opgenomen.

Hoofdstuk 4

Na te hebben vastgesteld wat beoordelaars zien (Hoofdstuk 2) en wat beoordelaars opschrijven (Hoofdstuk 3) met betrekking tot studentprestaties tijdens de stationstoets, trachtten we in Hoofdstuk 4 te onderzoeken hoe beoordelaars reeksen narratief commentaar interpreteren die uit een stationstoets verkregen zijn. Narratief beoordelingscommentaar over de communicatieve vaardigheden evenals globale communicatiescores werden verkregen voor 24 studenten in zes verschillende stations van een summatieve stationstoets. Over alle stations verzamelde narratieve gegevens werden geïncludeerd voor negen studenten (drie goede, drie gemiddelde en drie slechte presteerders gebaseerd op hun communicatiescore). Voor elk van de studenten namen tien expert-beoordelaars de verzamelde reeks narratief commentaar door. Cognitieve (informatie)verwerking werd vastgelegd door middel van hardopdenkprocedures en verbale protocolanalyse. We constateerden dat expert-beoordelaars hoofdzakelijk van twee strategieën gebruik maakten voor het interpreteren van narratief commentaar, namelijk: 1) vergelijken en contrasteren; en 2) het zich een voorstelling maken (visualiseren) van de studentprestaties. Bij het interpreteren van narratief beoordelingscommentaar bleken de beoordelaars echter drie verschillende perspectieven te gebruiken: 1) dat van de student (waarbij zij zichzelf in de schoenen van de student plaatsen); 2) dat van de examinator (waarbij zij de rol van de examinator overnemen en het geschreven commentaar volgens eigen normen of overtuigingen herinterpreteren); of 3) dat van de professional (waarbij zij het perspectief aannemen van de poortwachter van het beroep door de toets te beschouwen als een afspiegeling van de dagelijkse praktijk). Deze bevindingen benadrukken dat er bij de interpretatie van narratief beoordelingscommentaar sprake is van beoordelaarsvariantie die van invloed kan zijn op de manier waarop beoordelaars prestaties beoordelen en daarmee mogelijk ook op kwaliteit van de toets en de gegeven oordelen (bijv. betrouwbaarheid).

Hoofdstuk 5

In Hoofdstuk 5 onderzochten we de invloed van narratieve beoordelingsgegevens op de beoordelingskwaliteit door narratieve beoordelingsgegevens van een stationstoets te verkrijgen en de betrouwbaarheid van op narratief commentaar gebaseerde communicatiescores te vergelijken met die van uit directe observatie van studentgedrag verkregen scores. Narratief commentaar en communicatiescores (schaal 1-5) werden verkregen voor 14 afstuderende studenten farmacie over zes stations van de summatieve stationstoets, waarbij per station steeds twee beoordelaars de studentenprestaties direct observeerden en waardeerden (m.b.v. een score en narratief commentaar). Twee beoordelaars die de studenten niet geobserveerd hadden tijdens de toets, bekeken het narratief commentaar en beoordeelden op basis daarvan onafhankelijk van elkaar de communicatieve vaardigheden met een score, gebruik makend van dezelfde globale 5-puntsschaal die gebruikt werd tijdens de toets. Om de betrouwbaarheid (reproduceerbaarheid van toetsresultaten) te schatten werd gebruik gemaakt van generaliseerbaarheidstheorie. Om het verband tussen de scores van elke beoordelingsmethode te bepalen werden correlaties tussen de scores berekend. We ontdekten dat beide condities een vergelijkbare reproduceerbaarheid vertoonden. De generaliseerbaarheidscoëfficient bedroeg 0,571 voor de door de bij de stationstoets aanwezige beoordelaars aangeleverde scores en 0,612 voor de scores verschaft door beoordelaars die deze scores enkel op narratief commentaar gebaseerd hadden. Deze bevindingen maken aannemelijk dat de betrouwbaarheid van toetsscores voor communicatieve vaardigheden niet afhankelijk is van directe observatie tijdens de stationstoets. De correlatie tussen de twee scorereeksen bedroeg 0,5, hetgeen suggereert dat er in narratief commentaar andere aspecten van studentgedrag worden meegenomen/ meegewogen dan in de scores die verkregen werden tijdens de toets. Al met al maken deze bevindingen aannemelijk dat narratief commentaar van betekenis kan zijn binnen stationstoetsen en ondersteunen zij een verdere verkenning van de bruikbaarheid van narratief beoordelingscommentaar in de praktijk.

Hoofdstuk 6: Discussie

In Hoofdstuk 6 beantwoorden we onze onderzoeksvragen door de resultaten van onze studies samen te vatten en de betekenis ervan te interpreteren in relatie tot de bestaande literatuur. Ook reiken we handvatten aan voor de praktijk, doen we suggesties voor toekomstig onderzoek en bespreken we kort de sterke punten en beperkingen van dit proefschrift. Als antwoord op onze eerste onderzoeksvraag “Hoe verwerken beoordelaars prestatiegegevens bij het beoordelen van studenten op hun communicatieve vaardigheden tijdens de stationstoets?” kunnen we concluderen dat de wijze waarop beoordelaars informatie verwerken (d.w.z. waar zij op letten, hoe zij observaties vertalen naar narratieve gegevens en hoe zij narratieve prestatiegegevens interpreteren) van invloed kan zijn op de manier waarop zij een oordeel vormen over prestaties en/of deze scoren. Onze bevindingen tonen aan dat deze processen mogelijk door een aantal factoren beïnvloed worden, waaronder beoordelaars’ culturele oriëntaties en het perspectief vanuit welke zij de prestatiegegevens interpreteren. In antwoord op onze tweede onderzoeksvraag: “Hoe beïnvloedt het gebruik van narratieve beoordelingsgegevens de beoordelingskwaliteit?” kunnen we concluderen dat, ondanks verschillen in de wijze waarop beoordelaars narratieve gegevens interpreteren, de betrouwbaarheid van toetsresultaten (uitgedrukt als scores) niet afhankelijk is van directe observatie van studentprestaties. Daarbij dient te

worden opgemerkt dat het lijkt dat narratieve gegevens mogelijk op een iets andere manier de communicatieve vaardigheden tijdens de stationstoets meten dan scores op basis van directe observatie. Deze bevindingen ondersteunen een verdere verkenning van het gebruik van narratieve gegevens bij de stationstoets ter onderbouwing van beoordelaars' oordelen en ter verbetering van de geloofwaardigheid en verdedigbaarheid van de score/de beoordeling.

Valorisation

1. **(Relevance) What is the social relevance of your research results (i.e. in addition to the scientific relevance)?**

The results from this research can have potential social advantages for health professional training schools and the development of performance-based assessments.

As described in Chapter 2, we found that culture (as defined by Cultural Dimensions Theory) may influence how assessors interpret and judge communication behaviours. We found that assessors' cultural dimensions of masculinity-femininity and power-distance helped to explain differences in scoring, especially for borderline performers. This is an important finding in a world that is rapidly changing with migration and globalization. Although we are yet to fully understand the impact of culture on assessment, programs must be aware of potential cultural bias and work to ensure assessment practices are fair and provide equal opportunity for all students to be successful.

Our findings from Chapters 3 and 5 support the use of narrative assessment comments in Objective Structured Clinical Examinations (OSCEs) and this may improve defensibility of assessment decisions. In Chapter 3, we found that assessors write narrative comments that can distinguish between good and poor performers and focus on core constructs that are deemed to be fundamental to the patient care process (confidence, adaptability, patient safety, and professionalism). In Chapter 5, we found that scoring based on these comments was just as reliable as scores obtained during the interaction. Having written comments that are known to be reliable and discriminatory of student performance may supplement assessment decisions and provide a better understanding of what actually occurred during the OSCE. This 'record' of performance may therefore aid defensibility of OSCE outcomes, next to providing meaningful feedback to trainees.

Findings from the studies included in this thesis may have advantages with respect to assessor training. Findings from Chapters 2 and 4 showed that assessors' inherent characteristics (cultural dimensions, perspectives) influence how they approach and interpret performance data. The findings from our chapters support a more targeted approach to training by making assessors aware of their own 'approach' to assessment and have the capacity to self-reflect on their characteristics that may be influencing how they interpret and judge performance. If proven to be successful, shifting training accordingly may result in achieving better training outcomes.

2. (Target groups) To whom, in addition to the academic community, are your research results of interest and why?

Results of this thesis may be of interest to many groups and organizations. One major target would be teachers or educators within academic program committees that design performance assessments, including OSCEs. These committees may be interested in how obtaining narrative assessment data from OSCEs may support or enhance assessment approaches in which numeric scores are used to capture and make decisions about student performance. Additionally, our results from Chapters 2 and 4 add to previous research findings by showing that assessor characteristics are important for the interpretation and judgment of student performance in the standardized OSCE format. These results are of interest for faculty developers who must ensure that these issues are paid attention to in assessor training, as well as educators who are responsible for design of assessment systems, as our findings emphasize the need to include multiple assessors and multiple assessor perspectives in our decision making processes.

Secondly, students themselves may be a target group for the results of this research. Participant students were very interested in the narrative comments throughout the research and requested to obtain their own copies in some circumstances. Anecdotally, students seemed to think that comments would help them better understand assessors' scoring decisions and provide individualized performance information that would help to rationalize their grades. Although the perspectives of students were not sought from research within this thesis, it should be a major priority for future study.

Finally, organizations that implement OSCEs for professional licensure or renewal (e.g. Pharmacy Examining Board of Canada, Ontario College of Pharmacists, Pharmaceutical Society of New Zealand) may be interested in the results of this thesis. The findings relating to assessor characteristics, as well as the implications for assessor training and assessment instruments, may be advantageous for these organizations to consider when designing more robust assessment procedures.

3. **(Activities/Products) Into which concrete products, services, processes, activities or commercial activities will your results be translated and shaped?**

The findings from this thesis can inform the development of new assessor training programs for performance-based assessments of communication skills, including OSCEs. As discussed above, our findings support a targeted approach to assessor training that supports promotion of self-awareness and reflection by assessors, creating awareness of their inherent characteristics and beliefs that may influence the assessment process. Findings from Chapters 2 and 4, for example, suggest assessors could be trained to recognize and acknowledge their own cultural norms or assessment perspectives when observing and interpreting performance. If narratives are to be incorporated into OSCE assessment, as supported by our findings from Chapters 3 and 5, assessors will also need to be trained to write meaningful data within their narrative descriptions of performance. Our results are therefore very relevant to individuals or groups who administer OSCEs and recruit/train assessors.

Findings from this research are already shaping new OSCE assessment methods in different international contexts. The studies took place at Qatar University, where a final cumulative OSCE is maintained as an exit-from-degree requirement for students to complete prior to graduation. The results from Chapters 2 and 3 informed the development of a new communication assessment process, including the collection of narrative data. Having left Qatar and assuming a new post at the University of Otago in New Zealand, I am now implementing similar assessment approaches for our formative and summative OSCEs in my capacity as the Chair of the BPharm Undergraduate Curriculum Committee. These processes include modifications to our communication assessment instruments (as described in Chapter 3), as well as collection of narrative data. Dissemination and future testing of these new processes may result in uptake in other local and international settings.

Three studies from this thesis (Chapters 2, 3, and 5) are available as published manuscripts or available early online for purchase or academic use and have already received citations and over 120 reads on researchgate.net. Chapter 4 is currently under editorial review. The research has also gained interest at international conferences (the Netherlands, New Zealand, Qatar, Canada, Scotland, United Arab Emirates, and Finland), where I have spoken about our findings. . More speaking engagements are being scheduled in other countries in 2019, including Italy. These opportunities will help to disseminate and promote our findings to target groups and audiences.

4. **(Innovation) To what degree can your results be called innovative in respect to the existing range of products, services, processes, activities and commercial activities?**

To our knowledge, our results are the first to investigate the quality of narrative comments (Chapters 3 and 5) from OSCEs. OSCEs are traditionally scored based on checklists, rating scales, or rubrics and so our results provide an alternative approach to assessing student performance given that narrative assessment comments are reliable and can discriminate between students.

Another area where our results show innovation is the association between assessors' cultural dimensions and scoring of communication skills. To date, research has focused on the influence of culture on communication but yet to show any association with assessment of communication. Our results fill this gap by providing some evidence that assessors' culture may be a factor to consider when designing assessments and forming assessor groups or pairs. However, more research is required to better understand the relationship between culture and assessment and how cultural norms and preferences may influence performance judgements and decisions.

5. **(Schedule & Implementation) How will this/these plan(s) for valorization be shaped? What is the schedule, are there risks involved, what market opportunities are there and what are the costs involved?**

The valorization of this research can be divided into different streams. First, dissemination of the findings has already begun. To date, three of the chapters are publications available for purchase/downloading online and a fourth is under editorial review. I have also supported these publications with published commentary in journals and scholarly blogs to enhance dissemination of our findings. The work has also been presented at numerous local and international conferences, as described above. Links and key findings have also been shared on social media. The thesis will also be printed as a book and will be publicly available in 2019.

The second area of valorization is the practical uptake of our findings, including the use of narrative comments within OSCEs. As stated above, this is now occurring in two contexts (Qatar and New Zealand). As we gain more experience with using these comments for assessment purposes, we plan to further disseminate our work to the international community, which may increase uptake in other settings. A strong presence at international conferences and on social media platforms over the next 1-2 years should help to expedite this process and facilitate interest and uptake of our results.

Other areas of impact, such as the redesign of assessor training programs, may take longer (within 5 years) to produce. Before new training methods can be implemented, research and testing is required to ensure that training is effective and ultimately enhances the assessment process.

SHE Dissertations Series

SHE dissertations series

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to: https://she.mumc.maastrichtuniversity.nl

Van Rossum, T. (28-2-2019) Walking the tightrope of training and clinical service; The implementation of time variable medical training

Amalba, A. (20-12-2018) Influences of problem-based learning combined with community-based education and service as an integral part of the undergraduate curriculum on specialty and rural workplace choices

Melo, B. (12-12-2018) Simulation Design Matters; Improving Obstetrics Training Outcomes

Olmos-Vega, F. (07-12-2018) Workplace Learning through Interaction: using socio-cultural theory to study residency training

Chew, K. (06-12-2018) Evaluation of a metacognitive mnemonic to mitigate cognitive errors

Sukhera, J. (29-11-2018) Bias in the Mirror. Exploring Implicit Bias in Health Professions Education

Mogre, V. (07-11-2018) Nutrition care and its education: medical students' and doctors' perspectives

Ramani, S. (31-10-2018) Swinging the pendulum from recipes to relationships: enhancing impact of feedback through transformation of institutional culture

Winslade N. (23-10-2018) Community Pharmacists' quality-of-care metrics. A prescription for improvement

Eppich, W. (10-10-2018) Learning through Talk: The Role of Discourse in Medical Education

Wenrich, M. (12-09-2018) Guided Bedside Teaching for Early Learners: Benefits and Impact for Students and Clinical Teachers

Marei, H. (07-09-2018) Application of Virtual Patients in Undergraduate Dental Education

Waterval, D. (26-04-2018) Copy but not paste, an exploration of crossborder medical curriculum partnerships

Smirnova, A. (04-04-2018) Unpacking quality in residency training and health care delivery

Hikspoor, J. (05-12-2017) Development of the heart and vessels in the caudal part of the human body

Boymans, T. (06-10-2017) Hip arthroplasty in the very elderly: anatomical and clinical considerations

Zaidi, Z. (04-10-2017) Cultural hegemony in medical education: exploring the visibility of culture in health professions

Harrison, C. (20-09-2017) Feedback in the context of high-stakes assessment: can summative be formative?

Mekonen, H. (30-06-2017) Development of the axial musculo-skeletal system in humans

Taylor, T. (29-03-2017) Exploring Fatigue as a Social Construct: Implications for Work Hour Reform in Postgraduate Medical Education

McLellan, L. (29-03-2017) Prescribing the right medicine: Perspectives on education and practice

Ignacio, J. (09-02-2017) Stress Management in Crisis Event Simulations for Enhancing Performance

Bolink, S. (19-01-2017) Functional outcome assessment following total hip and knee arthroplasty; Implementing wearable motion sensors

Acknowledgements

Acknowledgements

It is difficult to believe that this process started over 5 years ago at a seaside restaurant over drinks in Doha, Qatar. I was new to academia and was trying to figure out a way to expand my knowledge and skills into new areas. Zubin Austin was visiting our university for OSCE training and as fellow Canadians, a group of us decided to take him out in a social setting. It was during this meeting that Zubin mentioned Maastricht University and the potential opportunity of pursuing a PhD in health professions education. Little did I know that this would become my reality within the following year.

Zubin's influence on my career is profound and I am thankful for his mentorship throughout the PhD process and helping me to develop expertise in OSCEs and performance-based assessments. It has been a true pleasure working with him and sharing many experiences that only living and working in Qatar can bring. I look forward to continued collaborations, as we strive to increase educational scholarship in the pharmacy profession across the globe.

I must say that I am very lucky to have been assigned my supervisors of Diana Dolmans and Marjan Govaerts – bringing two very different perspectives that challenged me to always better myself and to never settle for anything lower than perfection. They were patient when I first struggled with theory and always had a unique ability to help me clearly and concisely articulate my ideas and findings. I will greatly miss our monthly videoconferences, whether it be from Qatar, Canada, New Zealand, China, UK, or any other country we happened to be in at the time!

I want to extend my sincere appreciation to Dr. Ayman El-Kadi and Dr. Mohammad Diab from Qatar University who supported me throughout the last 5 years. Alongside these two, I would not have been able to complete this PhD without the engagement and interest of the amazing students and academic staff at the College of Pharmacy at Qatar University. Whether it was writing narratives, rating videos, or participating in training, this work would not have been possible without your support.

I want to thank Kerry Wilbur, who is my 'PhD partner in crime' and for making the dream of us drinking beer together in Maastricht come true. Special thanks go to Jennifer Morgan, who navigated me through my initial theoretical framework and was always there to listen to my updates over lunches in Saskatoon. I also want to thank Bridget Javed for her endless support during my time in Qatar and throughout the duration of the PhD.

Finally, I need to thank my family for their support and encouragement. My parents, John and Diane Wilby are forever by my side and will literally pick up and travel across the world to celebrate my accomplishments together. Last but definitely not least, I am forever grateful to my husband, Willian Amorim, who has now spent hours at home, in the car, in planes and airports, listening to me discuss my work, bounce ideas, and strive toward the finish line. Little did he know in 2016 that being in relationship with me also meant being in a relationship with my PhD. Te amo muito, HB!

Thank you, Dank je, Shukran, Obrigado!