Patient Education and Counseling 102 (2019) 1164–1169

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

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

Kyle John Wilby[a,*,1], Marjan J.B. Govaerts[b], Diana H.J.M. Dolmans[b], Zubin Austin[c], Cees van der Vleuten[b]

[a] College of Pharmacy, Qatar University, PO Box 2713, Doha, Qatar
[b] School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Universiteitssingel 60, 6229 ER Maastricht, Netherlands
[c] Leslie Dan Faculty of Pharmacy, University of Toronto, 144 College St., Toronto ON, M5S 3M2, Canada

ARTICLE INFO

Article history:
Received 16 September 2018
Received in revised form 20 December 2018
Accepted 24 January 2019

Keywords:
Assessment
Communication
Medical education
Pharmacy education

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.

1. 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

* Corresponding author at: PO Box 56, School of Pharmacy, University of Otago, 18 Frederick St, North Dunedin, Dunedin 9054, New Zealand.
E-mail address: kyle.wilby@otago.ac.nz (K.J. Wilby).
[1] Present address: School of Pharmacy, University of Otago, PO Box 56, 18 Frederick St, North Dunedin, Dunedin 9054, New Zealand.

https://doi.org/10.1016/j.pec.2019.01.018
0738-3991/

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

2. Methods

2.1. 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).

2.2. 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).

2.3. 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.

2.4. 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 min to write narrative assessment comments per student (8 min of

observation, 9 min 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 h. Once complete, the two scores obtained for each narrative were combined into a final composite (summed) score.

2.5. 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].

3. Results

3.1. 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. 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.

3.2. 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.

3.3. 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

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

Table 1
Summary of communication scores obtained from the OSCE and from scoring of narrative.

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.

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.

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.

3.4. 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 ($p < 0.05$).

4. Discussion and conclusion

4.1. 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.

4.2. 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 1500 phrases to read and interpret. Secondly, the procedures employed provided assessors with 17 min to write narratives, as it was anticipated that it would take longer than the actual station time of 8 min 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 min 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].

4.3. 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.

4.4. 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).

Source of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

All authors report no competing interests to declare.

Acknowledgements

The authors would like to acknowledge the assessors who wrote and evaluated narrative comments for the purposes of this study.

References

[1] R. Papp, I. Borbas, E. Dobos, M. Bredehorst, L. Jaruseviciene, T. Vehko, S. Balogh, Perceptions of quality in primary health care: perspectives of patients and professionals based on focus group discussions, BMC Fam. Pract. 15 (2014) 128.
[2] S.V. Doubova, F.C. Guanais, R. Perez-Cuevas, D. Canning, J. Macinko, M.R. Reich, 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. 31 (2016) 834–843.
[3] M.A. Stewart, Effective physician-patient communication and health outcomes: a review, Can. Med. Assoc. J. 152 (1995) 1423–1433.
[4] S.W. Mercer, M. Higgins, A.M. Bikker, B. Fitzpatrick, A. McConnachie, S.M. Lloyd, P. Little, G.C. Watt, General practitioners' empathy and health outcomes: a prospective observational study of consultations in areas of high and low deprivation, Ann. Fam. Med. 14 (2016) 117–124.
[5] J.M. Kelley, G. Kraft-Todd, L. Schapira, K. Kossowsky, H. Riess, The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials, PLoS One 9 (2014) e94207.
[6] E.S. Holmboe, J. Sherbino, D.M. Long, S.R. Swing, J.R. Frank, The role of assessment in competency-based medical education, Med. Teach. 32 (2010) 676–682.
[7] R.M. Epstein, Assessment in medical education, N. Eng. J. Med. 356 (2007) 387–396.

[8] W. Setyonugroho, K.M. Kennedy, T.J. Kropmans, Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: a systematic review, Patient Educ. Couns. 98 (2015) 1482–1491.
[9] B. Hodges, J. Turnbull, R. Cohen, A. Bienenstock, G. Norman, Evaluating communication skills in the OSCE format: reliability and generalizability, Med. Educ. 30 (1996) 38–43.
[10] M. Van Nuland, W. Van den Noortgate, C. Van der Vleuten, J. Goedhuys, Optimizing the utility of communication OSCEs: omit station-specific checklists and provide students with narrative feedback, Patient Educ. Couns. 88 (2012) 106–112.
[11] E. Driessen, V. van der Vleuten, L. Schuwirth, J. van Tartwijk, J. Vermunt, The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study, Med. Educ. 39 (2005) 214–220.
[12] J.L. Hanson, A.A. Rosenberg, J.L. Lane, Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States, Front. Psychol. 4 (2013) 668.
[13] S. Ginsburg, J. McIlroy, O. Oulanova, K. Eva, G. Regehr, Toward authentic clinical evaluation: pitfalls in the pursuit of competency, Acad. Med. 85 (2010) 780–786.
[14] A. Kuper, S. Reeves, M. Albert, B.D. Hodges, Assessment: do we need to broaden our methodological horizons? Med. Educ. 41 (2007) 1121–1123.
[15] C.P.M. Van der Vleuten, L.W.T. Schuwirth, Assessing professional competence: from methods to programmes, Med. Educ. 39 (2005) 309–317.
[16] S. Ginsburg, C.P.M. van der Vleuten, K.W. Eva, The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data, Acad. Med. 92 (2017) 1617–1621.
[17] J. Bartels, C.J. Mooney, R.T. Stone, Numerical versus narrative: a comparison between methods to measure medical student performance during clinical clerkships, Med. Teach. 39 (2017) 1154–1158.
[18] S.M. Downing, Reliability: on the reproducibility of assessment data, Med. Educ. 38 (2004) 1006–1112.
[19] R. Bloch, G. Norman, Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68, Med. Teach. 34 (2012) 960–992.
[20] Canadian Council for Accreditation of Pharmacy Programs, Accredited Programs, (2017) . (Accessed 22 July 2018) http://www.ccapp-accredit.ca.
[21] Association of Faculties of Pharmacy of Canada, Educational Outcomes for First Professional Degree Programs in Pharmacy in Canada – June 4, 2017, (2017) . (Accessed 22 July 2018) http://afpc.info/node/39.
[22] K.J. Wilby, E.K. Black, Z. Austin, B. Mukhalalati, S. Aboulsoud, S.I. Khalifa, Objective structured clinical examination for pharmacy students in Qatar: cultural and contextual barriers to assessment, East. Mediterr. Health J. 22 (2016) 251–257.
[23] A.H. Sobh, Z. Austin, M. MI Izham, M.I. Diab, K.J. Wilby, Application of a systematic approach to evaluating psychometric properties of a cumulative exit-from-degree objective structured clinical examination (OSCE), Curr. Pharm. Teach. Learn. 9 (2017) 1091–1098.
[24] G. Regehr, S. Ginsburg, J. Herold, R. Hatala, K. Eva, O. Oulanova, Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance, Acad. Med. 87 (2012) 419–429.
[25] L.Q. Munoz, C. O'Byrne, J. Pugsley, Z. Austin, Reliability, validity, and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy, Indian J. Pharm. Educ. Res. 5 (2005) 1–12.
[26] McMaster Education Research, Innovation & Theory, G_String, (2015) . (Accessed 22 July 2018) http://fhsperd.mcmaster.ca/g_string/.
[27] M.T. Brannick, H. Tugba Erol-Korkmaz, M. Prewett, A systematic review of the reliability of objective structured clinical examination scores, Med. Educ. 45 (2011) 1181–1189.
[28] M. Comert, J.M. Zill, E. Christalle, J. Dirmaier, M. Hartner, I. Scholl, Assessing communication skills of medical students in Objective Structured Clinical Examinations (OSCE) – a systematic review of rating scales, PLoS One 11 (2016) e0152717.
[29] S. Ginsburg, G. Regehr, L. Lingard, K. Eva, Reading between the lines: faculty interpretations of narrative evaluation comments, Med. Educ. 49 (2015) 296–306.
[30] E.S. Holmboe, J. Sherbino, D.M. Long, S.R. Swing, J.R. Frank, The role of assessment in competency-based medical education, Med. Teach. 32 (2010) 676–682.
[31] K.W. Eva, Cognitive influence on complex performance assessment: lessons from the interplay between medicine and psychology, J. Appl. Res. Mem. Cogn. 7 (2018) 177–188.
[32] M.J. Govaerts, C.P. van der Vleuten, L.W. Schuwirth, Optimising the reproducibility of a performance-based assessment test in midwifery education, Adv. Health Sci. Educ. Theory Pract. 7 (2002) 133–145.
[33] L.W.T. Schuwirth, C.P.M. van der Vleuten, Programmatic assessment: from assessment of learning to assessment for learning, Med. Teach. 33 (2011) 478–485.