Examining clinical skills and abilities in psychology – implementation and evaluation of an objective structured clinical examination in psychology

Anna E. Sundström and Camilla Hakelind

Abstract

Purpose – *Assessment of complex clinical skills and abilities is a challenge in mental health education. In the present study, an objective structured clinical examination (OSCE) was adapted to psychology and implemented in a Master in Psychology program. The purpose of the present study was to examine aspects of validity of this OSCE.*

Design/methodology/approach – *A total of 55 students enrolled in the Master in the Psychology program at Umeå University, Sweden, participated in two OSCE occasions. In addition to OSCE data, questionnaires were administered immediately after the OSCE to students (*n *= 18) and examiners (*n *= 13) to examine their perceptions of the OSCE.*

Findings – *The results provided support for different aspects of validity. The level of internal consistency was close to acceptable, and there was a good correspondence between global ratings and checklist scores for many stations. However, adding an additional category to the global rating scale and reviewing some of the station checklists might improve the assessment further. The present cut-score of the OSCE was comparable to a cut-score set by the borderline regression model. In general, students and examiners perceived the OSCE as a high-quality examination, although examiners raised some issues that could improve the OSCE further.*

Originality/value – *In conclusion, OSCE is a promising assessment in psychology, both from a psychometric perspective and from a test-taker and examiner perspective. The present study is an important contribution to the field as there are only a few examples where OSCE has been used in clinical psychology, and to the best of the authors' knowledge, this paper is the first to evaluate the validity of such an assessment.*

Keywords *Validity, OSCE, Clinical psychology education, Competence-based assessment*

Paper type *Case study*

Anna E. Sundström and Camilla Hakelind both are based at Department of Psychology, Umeå University, Umeå, Sweden.

Received 20 October 2021
Revised 7 March 2022
Accepted 8 March 2022

This study was funded by the Umeå School of Education, Umeå University, Sweden.

Introduction

Professional education within the mental health field, such as psychology, psychiatry, social work and mental health nursing, involves a competency-based framework with a requirement that schools assess student competence regarding the integration and application of competences in practice (Bogo *et al.*, 2011). In clinical psychology, competence-based assessment is increasingly important as practitioners are trained in the use of assessment skills and evidence-based treatments. Using standardized, reliable and valid methods for assessment of students' clinical competence is, therefore, an ongoing challenge in the psychology subject (Kaslow *et al.*, 2009).

For criterion-related educational programs, for example, clinical psychology programs, it is important that there is a correspondence between the learning objectives, the teaching and

assessment to ensure that the objectives are attained. This is referred to as *constructive alignment* (Biggs, 2003). When such programs are increasingly emphasizing clinical skills and abilities, including self-awareness, empathy and professional approach, traditional examination methods as e.g. written examinations and essays, need to be combined or replaced with assessment methods that in a reliable and valid manner can assess complex practical clinical skills and abilities (Norcini, 2007; Kaslow *et al.*, 2009).

In the medical and professional health fields, OSCEs have routinely been used to assess medical students' skills in e.g. interaction, diagnosis and intervention and are considered the golden standard of competence-based assessment (Khan *et al.*, 2013). There is a vast literature reporting on the quality of these examinations, providing empirical support for their reliability and validity (Daniels and Pugh, 2018; Hodges, 2006).

The OSCE format is based on the principles of objectivity and standardization and involves numerous time-limited "stations" typically with specially trained actors playing the role of a patient presenting with one or more clinical symptoms. Each station assesses essential elements of one or more competencies. During each station, which is typically limited to 10–15 min, the student's performance is observed by an independent rater who evaluates the student using a rating scale (commonly both a checklist score and a global rating of performance), generating the student's score for the station. Summing the student's scores across all the stations in the OSCE produces an overall exam score that is thought to represent the student's general clinical ability (Bogo *et al.*, 2011; Khan *et al.*, 2013).

Although the OSCE methodology has its origins in medicine (Harden *et al.*, 1975), it is now used for assessing educational outcomes in several different health professions such as nursing, physical therapy, dentistry as well as mental health nursing (Murcott and Clarke, 2016), psychiatry (Hodges, 2003) and social work (Bogo *et al.*, 2011). The adaptation of OSCE to the assessment of clinical skills in the mental health area is concerned with the challenges of assessing more general, less instrumental qualities such as the use of empathy, rapport and ethics (Hodges *et al.*, 1998) as well as interpersonal skills in engaging, understanding and supporting a patient or client and demonstrating different psychotherapeutic techniques (Murcott and Clarke, 2016). Therefore, the adaptation of OSCE to the mental health area and the evaluation of the functioning of the OSCE in these areas deserve attention. There are only a few previous examples of adapting and using OSCE in clinical psychology contexts (Sheen *et al.*, 2015; Johnson *et al.*, 2018; Yap *et al.*, 2012). These studies have examined students' and teachers' perceptions of the OSCE, and in general, the results indicate that students and examiners are positive about the use of OSCE for assessing clinical competence in psychology (Johnson *et al.*, 2018; Sheen *et al.*, 2015; Yap *et al.*, 2012). However, examining students' and examiners' perceptions of the OSCE is only one aspect of the functioning of the examination. To our knowledge, no previous study has reported on the validity of an OSCE in psychology. In this paper, we report on the implementation of an OSCE in a clinical psychology program in Sweden and examine aspects of the validity of the OSCE through analyses of performance data from the OSCE as well as student and examiner evaluations of the OSCE.

Umeĺ University objective structured clinical examination adapted for the Master in Psychology program

The requirement for becoming a licensed psychologist in Sweden is 300 ECTS [1] credits/five-year master-level education in psychology followed by one year of supervised practice. During a revision of the Master in the Psychology program at Umeå University, Sweden, the OSCE format was adapted to clinical psychology and implemented in 2018 to enhance the alignment between the national goals and the examination of clinical skills and abilities. The OSCE is a mid-way assessment that is implemented in the sixth semester, shortly before the internship and also before the students start the psychotherapy training course where they, as student therapists, provide treatment for real clients. The OSCE is a

part of a course of 3.5 ECTS credits named the *Integrative course*. Preceding the integrative course, students take a number of clinical courses. The elements of the integrative course then provide an opportunity to practice and integrate the knowledge, skills and abilities taught in those previous courses, targeting those considered highly important in the psychologist profession. The pedagogic setup of the course is based on workshops, seminars, lectures and supervision. A large part of the pedagogical approach also involves group work where students practice their psychological skills through role-playing. The course ends with the OSCE, which is conducted for two days.

In the OSCE, each student undergoes a circuit of 10–12 stations where he or she encounters various problems/tasks that are intended to reflect several of the goals of clinical competences inherent in the curricula and national goals for the clinical psychologists' exam. The students start at different stations in the circuit to enable the whole group can start and finish the OSCE at the same time. The skills and abilities that are assessed in the OSCE focus on cognitive behavior therapy (CBT), psychodynamic therapy (PDT), motivational interviewing (MI), giving psycho-educative information, performing testing, diagnostic assessment and psychiatric evaluation. Several of the different stations contain more general criteria, which focus on the ability to display and express empathy, the ability to lead and maintain the focus of the session, as well as the ability to engage with and collaborate with one or two clients. At each station, students interact with a standardized patient (SP) played by an actor, and an examiner makes a structured assessment of the student's skills and abilities using a detailed checklist and a global rating of performance.

Indicators of quality of the objective structured clinical examination

An essential condition to guarantee high-quality and effective OSCE is to provide evidence that supports the validity of its scores (Hodges, 2003). The validity of a test is traditionally defined as the degree to which the test measures what it is intended to measure, and hence, the validity of a test should be accumulated by collecting several sources of evidence. The standards for educational and psychological testing (AERA *et al.*, 2014) describe five sources of validity evidence: content, response processes, internal structure, relations with other variables and consequences.

Evidence for *content validity* of the OSCE was ensured by the process of developing the station scenarios and checklists. The OSCE stations were blueprinted to match the clinical content in which the students were trained during the previous courses in the program as well as the integrative course, which were considered relevant to the psychologist profession and in accordance with the national goals. The teachers appointed to be examiners for each OSCE station were first provided with instructions about the OSCE method, after which they were asked to create case scenarios for the stations as well as clients' scripts, instructions to students and item checklists. All material for the stations was then reviewed and scrutinized in teacher teams of three to four teachers. The stations were pilot tested with the help of voluntary students and actors who played SPs. After the pilot testing, additional revisions of the instructions to students, case scenarios and checklists were made before the implementation of the first OSCE.

Validity evidence related to *response processes* ensures the accuracy of the data collected using the checklists and global ratings. Examiners used checklists and global ratings to assess student performance at each station. Checklists included about 8–10 items per station, and each checklist item was scored using two or three anchors (0–1 or 0–2 points). The global rating scale comprising four levels; excellent, clear pass, borderline pass and clear fail. To prepare for their role as patients, SPs were provided with manuscripts, including extensive descriptions of each scenario and instructions on how to act, approximately one month before the OSCE. Before starting the OSCE, the SP and the examiner were allowed to discuss the case and the acting instructions, and the examiner was given an opportunity to point out vital parts that need to be in place to allow students to

show their skills. Evidence based on response processes can also include information from test-takers and examiners. In the present study, students' and examiners' perceptions of the quality and relevance of the examination were examined through a Web-based questionnaire.

The *internal structure* validity evidence refers to the reliability and psychometric properties of the OSCE. The analyses used for collecting internal structure validity evidence in the present study are based on suggestions by Pell *et al.* (2010) and are described in the method section.

The appropriateness of the cut scores for the OSCE is closely linked to the *consequential aspect of validity* (Pant *et al.*, 2009; Yousuf *et al.*, 2015). There are several models that can be used for setting standards. The borderline-regression-model (BRM) is an examinee-centered standard-setting procedure that has been shown to provide reliable, credible and feasible standards for small-scale OSCEs (Kaufman *et al.*, 2000; Kramer *et al.*, 2003; Reid and Dodds, 2014; Yousuf *et al.*, 2015). In the present study, we used the BRM to set a pass-score for the OSCE and compared it with the current standard set at 60%.

Aim

The overall aim of the study was to describe the implementation of an OSCE in a Master in Psychology program and evaluate aspects of quality in terms of reliability and validity using performance data from the OSCE. In addition, the aim was to examine students' and examiners' perceptions of the quality of the OSCE.

Method

Procedure

A summative psychology OSCE was administered to students in the sixth semester of the Master in Psychology program at Umeå University, Sweden, in January 2020 and September 2020, respectively. The scenarios or "stations" in the OSCE were designed to represent authentic situations for a professional clinical psychologist. All stations involved encounters with (SPs). There were also two rest stations in each OSCE. At the rest stations, no SP or examiner was present and the student was given 10–15 min to relax. At each station, one examiner viewed students' interacting with the SP for 10 min and assessed student performance by completing a checklist (with a maximum score of 10). The checklists contained task-related criteria as well as more general criteria, which focused on the ability to display and express empathy, the ability to lead and maintain the structure of conversations, as well as ability to engage with and collaborate with one or two clients. At the end of each encounter, the examiner made a global rating of the student's performance.

The OSCE consisted of 12 and 10 stations, respectively, in the spring and autumn semester of 2020, depending on the number of examiners available. Due to the differences in the number of stations, the maximum station score in the spring semester OSCE was 120 points and 100 points in the autumn semester. The cut-score was set at 60%, which is a faculty-wide standard. In the present study, an alternate cut-score was established using the BRM, and this cut-score was compared to the current standard.

Participants

In total, 67 students enrolled in the Master in Psychology program took part in the OSCE in 2020. Participation in the OSCE examination was compulsory, but participation in the research project was voluntary. There were 55 students (34 females and 21 males) that agreed to participate in the study. Of those, 29 (19 female, 10 male) took the OSCE held in the spring semester of 2020 and 26 students in the autumn semester (15 female, 11 male). Out of the 26 students taking the OSCE in the autumn semester, 18 (69%) agreed to answer

the questionnaire. Of these, 11 were females and 7 males. All of the 13 examiners agreed to answer the examiner questionnaire. All participants gave their informed consent to participate in the study. Ethical approval was obtained from the Swedish Ethical Review Authority (Dnr 2020–01440).

Instruments

Student and examiner feedback on the OSCE were collected through Web-based questionnaires sent out to those participating in the autumn-semester OSCE immediately after the OSCE. The questionnaires were based on the ones originally developed and used by Yap *et al.*'s (2012) and later modified by Sheen *et al.* (2015), and permission to use the questionnaires was obtained. The student and examiner questionnaires included questions about different aspects of the perceived quality and relevance of the examination. The responses were rated on a 5-point Likert scale, ranging from 1= Strongly disagree to 5 = Strongly agree. The examiner questionnaire also included one open-ended question, where examiners were asked to comment on how they perceived the assessment of the students during the OSCE and whether they believed there were some problems or aspects that needed to be improved. The examiners' comments were categorized into themes and reported in the results section.

Statistical analysis

During the course of the OSCE, examiners entered checklist scores and global ratings via computer software developed especially for the OSCE at the Department of Psychology, Umeå University. Data was exported to Excel from the software, and statistical analyses were carried out using SPSS 25. Different psychometric analyses were performed to collect evidence for internal structure validity. The difficulty of the OSCE examination and the station categories were examined by overall examination mean score and standard deviation as well as the mean and standard deviation for each station category. Following the suggestions by Pell *et al.* (2010), we calculated Cronbach's alpha coefficient for the total OSCE checklist scores to assess the generalizability of performance scores across the OSCE stations. In addition, the relationship between checklist scores and global ratings was examined using the R^2 coefficient, which is the proportional change in the checklist score due to a change in the global assessment. A Cronbach's alpha of 0.70 or above and an R^2-value of 0.50 and above indicate acceptable reliability and a reasonable relationship between checklist scores and global grades (Pell *et al.*, 2010). Moreover, pass rates were examined, and the BRM was used to set the cut scores for the OSCEs. The BRM uses a linear regression approach, where checklist scores are regressed onto global ratings to set cut scores (Kramer *et al.*, 2003; Woehr *et al.*, 1991). By inserting the midpoint of the global rating scale corresponding to the borderline group into the equation, a corresponding predicted checklist score can be determined. This predicted score becomes the cut-score for the station. The sum of the station cut-scores constitutes the cut-score for the OSCE. The pass-score received from the BRM was compared to the current method for setting pass-scores for the OSCE, which is 60% correct of the checklist scores at the stations. Students with less than 60% correct of the total OSCE station scores or a global rating of "clear fail" at two or more stations fail the OSCE.

Results

In general, students performed well at the OSCE. The pass-rate was 100% in both semesters. The average total score for the OSCE held in the spring was 92 (out of 120; corresponding to an average of 76% correct). The corresponding figures for the OSCE held in the autumn were 73 (out of 100; corresponding to 73% correct) (Table 1).

Table 1 Descriptive statistics and internal consistency (alpha) for total OSCE checklist scores

		% Correct				
Occasion	n	*Min*	*Max*	*M*	*SD*	*α checklist score*
OSCE spring	29	60	86	76	0.06	0.66
OSCE autumn	26	60	86	73	0.05	0.45

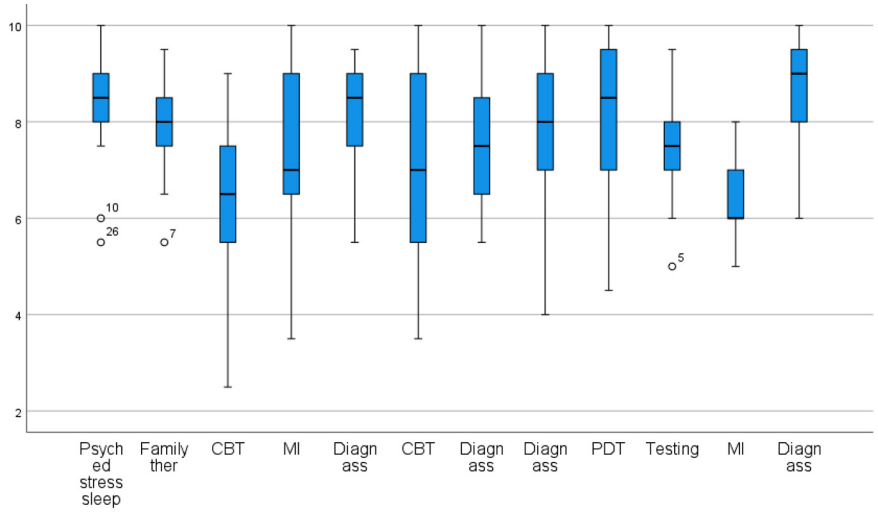
Note: Spring 2020: 12 stations; Autumn 2020: 10 stations

The OSCE stations were divided into seven different categories, and the distribution of checklist scores for each station was examined. Figures 1 and 2 show that the score distribution for many of the stations both in the spring and autumn semesters was restricted.

The difficulty of the stations within each category was examined by calculating the average checklist score (Table 2). The results indicate that the most difficult were the "CBT treatment" and "motivational interviewing" stations, and among the easiest stations were "providing psycho-educative information" and "diagnostic assessment" stations. The relative difficulty of the stations was similar between the spring and autumn semester, except for the PDT stations, which were among the easiest in spring 2020 and among the most difficult in the autumn.

As reported in Table 1, Cronbach's alpha for spring OSCE was 0.66 and for autumn OSCE 0.45. This suggests some evidence for generalizability of individual stations across the range of scenarios tested but indicates that the decrease from 12 to 10 stations had a negative impact on the reliability. The relationship between checklist scores and global ratings for each station was examined, and the results indicated that five stations during the spring semester OSCE and six stations during the autumn semester had values less than the recommended value of R^2>0.50. These stations were examined graphically by plotting checklist scores against global grades. For two stations (psycho-education about stress and sleep and family therapy – leading a session with two clients) in the spring OSCE the plots indicated that there was a rather large distribution of checklist scores among those who received the global grade "clear pass", and for one station (diagnostic assessment

Figure 1 Distribution of test score by station categories for the spring semester OSCE



Notes: Station categories are psychoeducation stress and sleep; family therapy; cognitive behavior therapy (CBT); motivational interviewing (MI); diagnostic assessment; psychodynamic therapy (PDT)

Figure 2 Distribution of test scores by station for the autumn semester OSCE

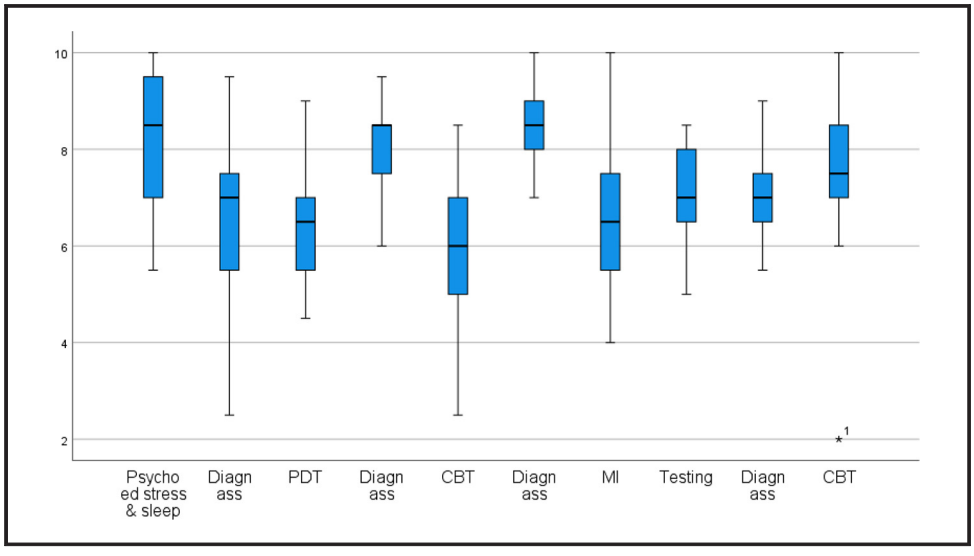


Table 2 Descriptive statistics for the difficulty of OSCE stations, *n* of stations, min-max, M and SD for checklist scores

	Spring					*Autumn*				
Station categories	n	*Min*	*Max*	*M*	*SD*	n	*Min*	*Max*	*M*	*SD*
Testing	1	5.0	9.5	7.40	0.93	1	6.0	8.50	7.39	0.90
Diagnostic assessment	4	6.13	9.63	8.08	0.87	4	6.63	9.50	7.58	0.57
CBT treatment	2	4.00	8.50	6.67	1.18	2	4.50	8.25	6.90	0.78
PDT treatment	1	4.5	10.0	8.17	1.62	1	4.5	9.00	6.35	1.02
Psychoeducative information	1	5.5	10.0	8.57	1.03	1	5.5	10.0	8.37	1.37
Motivational interviewing	2	4.75	9.00	6.92	1.06	1	4.0	9.50	6.48	1.35
Family therapy	1	5.5	9.5	7.95	0.94	0	–	–	–	–

Note: Maximum checklist score is 10

station – conveying results from a developmental evaluation) there were a few persons with low global grades (clear fail and borderline pass) that received relatively high checklist scores. For the two remaining stations with low R^2 values as well as for the stations in the autumn OSCE, the weak relationship could be explained by the restriction of range in global ratings and checklist scores.

Standard-setting using the borderline-regression-model

In this study, the BRM was applied to set the standard for the OSCE. For the spring semester, the cut score resulting from the BRM was 57% (score of 68) [2]. For the autumn semester OSCE, the BRM cut score was 55% correct (score of 55). This can be compared to the cut score of 60% that was used both semesters.

Students' and examiners' perceptions of the objective structured clinical examination

The questionnaire responses from the students and examiners indicated that they, in general, were fairly satisfied with the OSCE (Figures 3 and 4). A majority of the students and examiners responded that stations reflected topics in earlier teaching and that the content

Figure 3 Descriptive statistics (percentage) from the student questionnaire ($n = 18$)

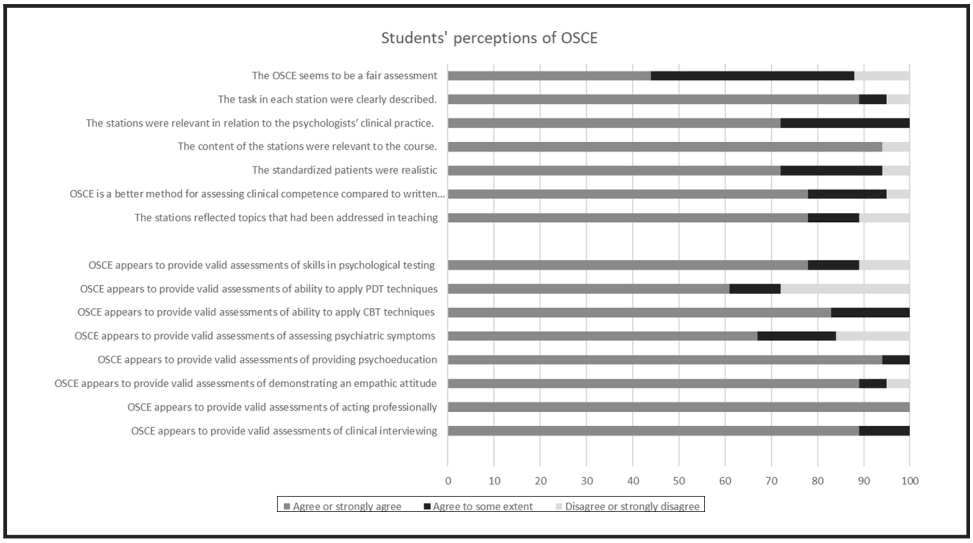
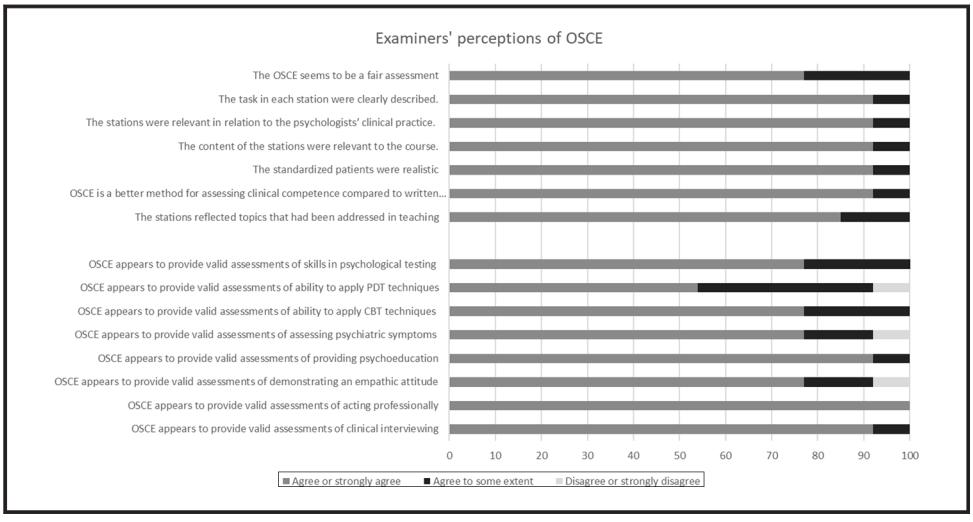


Figure 4 Descriptive statistics (percentage) from the examiner questionnaire ($n = 13$)



of OSCE was relevant for the course as well as for the role of the clinical psychologist. A majority also agreed that the OSCE is a better method for examining clinical competence than a written examination. Both students and examiners agreed that the stations were clearly described and that the SP's were realistic. Regarding the question of whether the OSCE is a fair assessment, most examiners agreed. However, the students were somewhat more uncertain about this.

The examiners and students were asked whether the OSCE appears to provide valid assessments of different aspects of clinical skills. Both students and examiners agreed that OSCE seems to provide valid assessments of clinical interviewing, acting professionally, demonstrating an empathic attitude and providing psycho-education. For applying PDT techniques and assessing psychiatric symptoms, both students and examiners were a bit more uncertain (Figures 3 and 4).

Examiners were asked to comment on their experiences of assessing the students at their OSCE station and on whether they thought there were problems or aspects that needed to be improved to make valid and reliable assessments in the OSCE. Their answers were categorized into three different themes; Developing and pilot-testing stations, assessment during the OSCE and feedback to students.

Developing and pilot-testing stations. Examiners commented on the checklists and global rating scale, describing it as challenging to develop stations/scenarios with a reasonable scope and to make the checklist criteria exhaustive and corresponding with the global assessment. Moreover, they thought that the case descriptions could be made even clearer for some stations. Examiners also commented that it would be valuable with a consensus discussion of the categories on the global rating scale among the examiners before each OSCE occasion. Furthermore, examiners believed that it is important to pilot-test new stations to evaluate the checklist and time frame, as well as to provide SP's with more training before the OSCE.

Assessment during the objective structured clinical examination. In general, the examiners thought the OSCE format worked well, that they received much information about student skills and abilities during the course of a station. However, they raised issues about the assessment, for example, that as an examiner, one gets tired during a whole day of examination and that this might affect the assessments.

Feedback. Finally, the examiners stressed the importance of the feedback to students based on their OSCE performance and that this feedback is valuable for students in their further training during the internship and the practice as student-therapists at the psychotherapy training course later-on in the program.

Discussion

Finding valid and reliable ways of assessing complex clinical skills is one of the major challenges in mental health education, such as in clinical psychologist programs (Stevens *et al.*, 2017). In other health professions, OSCE is a commonly used tool for the assessment of clinical competence. However, OSCE has rarely been applied to clinical psychology, and the few previous studies reporting on OSCE in psychology have focused on students' and examiners' perceptions of OSCE (Yap, 2012; Sheen *et al.*, 2015; Johnson *et al.*, 2018). To establish the quality of an OSCE, evidence is needed to verify the validity of the scores.

This study reported on the implementation of an OSCE in a Master in Psychology program and evaluated aspects of the reliability and validity related to this OSCE. Validity evidence related to content was collected through the process of developing and blueprinting the OSCE. Before the OSCE was implemented for the first time, a rigorous procedure of developing, scrutinizing and pilot-testing stations was conducted. However, the pilot-testing of stations was dropped for later OSCE occasions, which was commented upon by examiners, who thought pilot-testing and rater training would be needed before each OSCE occasion.

Internal structure validity evidence involves the analysis of psychometric properties of the OSCE. The average total checklist score on the OSCE in spring and fall, respectively, indicated that the two OSCE occasions were rather parallel in difficulty. Moreover, the results provided some evidence for the generalizability of individual station scores across the range of scenarios tested. In the spring OSCE that comprised 12 stations, Cronbach's alpha was 0.66. However, in the autumn OSCE, where the number of stations was decreased to 10, the Cronbach's alpha was lower; = 0.45. This indicates that a decrease from 12 to 10 stations had a negative impact on the reliability of the assessment and that a 12 station OSCE is preferable. Compared with OSCEs in other mental health areas, however, the results of our study are in line with reliabilities

reported in psychiatry (alpha = 0.51; Hodges, 1998) and social work (alpha = 0.55; Bogo *et al*., 2011). Provided that the stations in the OSCE assess different aspects of clinical competence in psychology and that they differ in difficulty level, a very high Cronbach's alpha is not to be expected.

The distribution of checklist scores over stations showed that scores were restricted in range for many of the stations. This was also the case for the global ratings. The stations in which the range of scores was most restricted need to be scrutinized to examine whether a revision of the checklist items can increase the discrimination between students at different proficiency levels, especially around the cut-score. This is important because one of the goals of the OSCE is to ensure the competence level of the students before moving on to internship and psychotherapy training courses to meet real clients. Comparing the different station categories, there was some variation in difficulty level between stations. Among the more difficult stations were the "CBT treatment" and "motivational interviewing" stations, and among the easiest stations were "providing psycho-educative information" and "diagnostic assessment" stations. The relative difficulty of the stations was similar between the spring and autumn semester, except for the PDT stations, which were among the easiest in spring 2020 and among the most difficult in the autumn. In these stations, there was a change of examiner between the spring and the autumn, which might be one explanation for this difference. This indicates the importance of training the raters to calibrate their assessments within and over semesters.

The correspondence between checklist scores and global ratings for each station was examined by the R^2 coefficient, and the results showed that three stations had a mismatch between checklist scores and global ratings. These results imply that checklists for some stations can be a poor indicator of ability. Thus, a revision of checklist items for these stations to optimize correspondence with global rating categories would be recommended, for example, by ensuring that checklist items have three instead of two anchors where appropriate, thereby allowing greater discrimination. The fact that many stations had restricted range in global ratings also indicates that adding an extra category, "borderline fail" on the global rating scale to increase the discrimination between students around the cut-score would improve the assessment in terms of providing more information about student performance. Such a revision would also make the global rating scale more balanced.

Choosing a defensible passing score by using a suitable standard-setting method is a key issue for validity (Pant *et al*., 2009). The standard-setting procedure is of great importance for a consequential aspect of validity as the choice of passing score has a direct impact on the decisions made based on the test score and its consequences for the individual test-taker. BRM is a standard-setting model that has been suggested for small-scale OSCEs because it uses data from all examinees in calculating the pass-score instead of only the borderline group (Pell *et al*., 2010). In this study, the established cut-score of 60% of the OSCE station checklist score was compared to the pass-score calculated with BRM. For both semesters, the cut-off set by the BRM was lower than the currently used 60% pass-score. For the OSCE rounds in the spring semester, the BRM cut-score was 57% and for the OSCE rounds in the autumn, 55%. Yet, with both methods, the pass-rate in the student sample was 100%, which deserves comment. In general, this student population is high performing, as indicated by their performance on other examinations during the program. However, previous and later OSCE occasions using similar stations and the same pass-score have resulted in at least a couple of students failing the examination. The lack of variability in the global rating scale and checklists for some of the stations indicates that the scales might need a revision to increase the discrimination between students at different proficiency levels, for example, by adding a category on the global rating scale. Such revisions would require a re-calculation of the cut-score. From this perspective, it is also possible that the current cut-score might result in some false positives, i.e. that students just

above the borderline limit would fail the OSCE if more detailed and discriminating checklists and global rating scales were used. Therefore, any revision of the global rating scale or the checklist should ideally be followed up with pilot testing, ensuring that the revisions result in the desired quality improvement. An additional way of increasing correspondence between the global rating and the checklist is by establishing consensus regarding the definition of borderline performance. This could be done by letting a number of examiners watch previous OSCE station recordings with students who exhibit borderline performance and discuss the criteria for the assessment. Such a procedure could also help improve inter-rater reliability among examiners.

Validity evidence related to the response processes was provided through the training of the examiners, the development of checklists and the training of the SPs. Moreover, evidence of the response processes was also indicated by students' and examiners' perceptions of the OSCE collected in the questionnaire (AERA *et al.*, 2014). Examining students' perceptions of and attitudes toward OSCE is a common feature in studies evaluating OSCEs in other subjects (Jay, 2007; Larsen and Jeppe-Jensen, 2008) as well as in the few previous studies of OSCEs in psychology (Sheen *et al.*, 2015; Yap *et al.*, 2012). Some studies have also included examiners' perceptions (Johnson *et al.*, 2018; Sheen *et al.*, 2015). The results from the present study revealed that most of the students and examiners had a positive perception of the OSCE in terms of clarity of instructions, the authenticity of SP's, relevance of the station content, as well as to the role of the clinical psychologist. Regarding the question of whether OSCE is viewed as a fair form of assessment, a majority of the examiners agreed, whereas students were somewhat more doubtful. Furthermore, although a small number of students and examiners were unsure, most agreed that the OSCE was a valid method for assessing clinical competence. These results are in line with previous studies that have examined students' and examiners' perceptions of OSCE in psychology (Johnson *et al.*, 2018; Sheen *et al.*, 2015; Yap *et al.*, 2012), indicating that students and examiners overall view OSCE as a valid, realistic and fair assessment method.

Moreover, the examiner responses indicated that they thought the feedback from the OSCE would be valuable for students in their further training and development of clinical skills. Examiners raised some issues about the difficulty of develop stations of reasonable scope, exhaustive checklist criteria and achieve correspondence between checklists and global ratings. In addition, they believed that it would be a good idea that examiners to discuss the meaning of the global rating categories before each OSCE occasion. Taken together with the results from the psychometric analyses, this indicates the need for continuous training of examiners to calibrate their assessments, as well as refining the checklists and improving the discrimination of the global rating scale by adding a category to improve measurement properties.

Limitations and further studies

This study was based on data from the OSCE in psychology for two semesters. Because only students in the sixth semester take the OSCE, the sample used in the present study was relatively small. On the other hand, this study makes an important contribution to the field because it is the first to our knowledge and reporting on the implementation of an OSCE in clinical psychology and evaluating the validity of the assessment. In the present study, reliability was examined in terms of internal consistency. Another important aspect of reliability for performance assessments such as OSCES is, of course, inter-rater reliability (Pell *et al.*, 2010); because the present study only involved one rater per station, examination of inter-rater reliability was not possible. Further studies of the OSCE in clinical psychology should include inter-rater reliability to examine the agreement between examiners, which can also guide rater training and calibrate raters for future OSCE occasions. In future studies, it would also be desirable to include validity evidence of

relationships to external variables (AERA *et al.*, 2014), e.g. to examine the correlation between OSCE performance and clinical placement supervisor competency ratings during an internship.

Conclusions

The implementation of OSCE in psychology at the Department of Psychology, Umeå University, Sweden, provides evidence for the reliability and validity of this tool in assessing psychology students' clinical skills and abilities. In general, the findings provided support for validity related to the content, response processes, internal structure and consequences of the assessment. The results indicated that a 12 station OSCE is preferred because it provides a more reliable assessment than a 10 station OSCE. The distribution of global ratings and checklist scores and the relationship between these indicated that refining checklists as well as adding a category on the global rating scale to increase discrimination might improve measurement properties of the OSCE further. In general, both examiners and students believed the OSCE was a valid assessment of clinical competence, although some issues were raised by examiners. For example, supporting examiners in the development of exhaustive checklists, providing training for new examiners and arranging a consensus discussion of the global rating scale before each OSCE occasion was suggested. To conclude, the present study indicates that OSCE is a useful form of assessment of clinical competence in psychology, both from a psychometric perspective and from a test-taker and examiner perspective. However, there is, as always, room for improvement.

Notes

1. European Credit Transfer System.
2. For two of the stations in the spring semester (Stations 11 and 12), there were no variation in the global ratings, resulting that no linear regression could be calculated. The cut-score for these specific stations were replaced with the average cut-score for the other stations.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (Eds) (2014), *Standards for Educational and Psychological Testing*, American Educational Research Association, Lanham, MD.

Biggs, J. (2003), "Teaching for quality learning at university", The Society for Research into Higher Education & Open University Press, Glasgow.

Bogo, M., Regehr, C., Logie, C., Katz, E., Mylopoulos, M. and Regehr, G. (2011), "Adapting objective structured clinical examinations to assess social work students' performance and reflections", *Journal of Social Work Education*, Vol. 47 No. 1, pp. 5-18, doi: 10.5175/JSWE.2011.200900036.

Daniels, V.J. and Pugh, D. (2018), "Twelve tips for developing an OSCE that measures what you want", *Medical Teacher*, Vol. 40 No. 12, pp. 1208-1213, available at: www.tandfonline.com/doi/pdf/10.1080/0142159X.2017.1390214?needAccess=true

Harden, R., Stevenson, M., Downie, W. and Wilson, G. (1975), "Assessment of clinical competence using objective structured examination", *BMJ*, Vol. 1 No. 5955, pp. 447-451.

Hodges, B. (2003), "Validity and the OSCE", *Med Teach*, Vol. 25 No. 3, pp. 250-254, doi: 10.1080/01421590310001002836.

Hodges, B., Regehr, G., Hanson, M. and McNaughton, N. (1998), "Validation of an objective structured clinical examination in psychiatry", *Acad Med*, Vol. 73 No. 8, pp. 910-912, doi: 10.1097/00001888-199808000-00019.

Hodges, B.D. (2006), "The objective structured clinical examination: three decades of development", *Journal of Veterinary Medical Education*, Vol. 33 No. 4, pp. 571-577, doi: 10.3138/jvme.33.4.571.

Jay, A. (2007), "Students' perceptions of the OSCE: a valid assessment tool?", *British Journal of Midwifery*, Vol. 15 No. 1, pp. 32-37, doi: 10.12968/bjom.2007.15.1.22677.

Johnson, H., Mastroyannopoulou, K., Beeson, E., Fisher, P. and Ononaiye, M. (2018), "An evaluation of multi-station objective structured clinical examination (OSCE) in clinical psychology training", *Clinical Psychology Forum*, Vol. 301, pp. 38-43.

Kaslow, N.J., Grus, C.L., Campbell, L.F., Fouad, N.A., Hatcher, R.L. and Rodolfa, E.R. (2009), "Competency assessment toolkit for professional psychology", *Training and Education in Professional Psychology*, Vol. 3 No. 4, Suppl, pp. S27-S45, doi: 10.1037/a0015833.

Kaufman, D.M., Mann, K.V., Muijtjens, A.M.M. and van der Vleuten, C.P.M. (2000), "A comparison of standard-setting procedures for an OSCE in undergraduate medical education", *Academic Medicine: Journal of the Association of American Medical Colleges*, Vol. 75 No. 3, pp. 267-271.

Khan, K.Z., Ramachandran, S., Gaunt, K. and Pushkar, P. (2013), "The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective", *Medical Teacher*, Vol. 35 No. 9, pp. e1437-e1446, doi: 10.3109/0142159X.2013.818634.

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L. and Van Der Vleuten, C. (2003), "Comparison of a rational and an empirical standard setting procedure for an OSCE", *Medical Education*, Vol. 37 No. 2, pp. 132-139, doi: 10.1046/j.1365-2923.2003.01429.x.

Larsen, T. and Jeppe-Jensen, D. (2008), "The introduction and perception of an OSCE with an element of self- and peer-assessment", *European Journal of Dental Education*, Vol. 12 No. 1, pp. 2-7, doi: 10.1111/j.1600-0579.2007.00449.x.

Murcott, W.J. and Clarke, N. (2017), "Objective structured clinical exam: a successful approach to pre-registration mental health nurse assessment", *The Journal of Mental Health Training, Education and Practice*, Vol. 12 No. 2, pp. 90-97, doi: 10.1108/JMHTEP-06-2016-0031.

Norcini, J. (2007), *Workplace-Based Assessment in Clinical Training*, Association for the Study of Medical Education, Edinburgh.

Pant, H.A., Rupp, A.A., Tiffin-Richards, S.P. and Köller, O. (2009), "Validity issues in standard-setting studies", *Studies in Educational Evaluation*, Vol. 35 Nos 2/3, pp. 95-101, doi: 10.1016/j.stueduc.2009.10.008.

Pell, G., Fuller, R., Homer, M. and Roberts, T. (2010), "How to measure the quality of the OSCE: a review of metrics – AMEE guide no. 49", *Medical Teacher*, Vol. 32 No. 10, pp. 802-811, doi: 10.3109/0142159x.2010.507716.

Reid, J.K. and Dodds, A. (2014), "Comparing the borderline group and borderline regression approaches to setting objective structured clinical examination cut scores", *Journal of Contemporary Medical Education*, Vol. 2 No. 1, pp. 8-12.

Sheen, J., McGillivray, J., Gurtman, C. and Boyd, L. (2015), "Assessing the clinical competence of psychology students through objective structured clinical examinations (OSCEs): student and staff views", *Australian Psychologist*, Vol. 50 No. 1, pp. 51-59, doi: 10.1111/ap.12086.

Stevens, B., Hyde, J., Knight, R., Shires, A. and Alexander, R. (2017), "Competency-based training and assessment in Australian postgraduate clinical psychology education", *Clinical Psychologist*, Vol. 21 No. 3, pp. 174-185, doi: 10.1111/cp.12061.

Woehr, D.J., Arthur, W. and Fehrmann, M.L. (1991), "An empirical comparison of cutoff score methods for content-related and criterion-related validity settings", *Educational and Psychological Measurement*, Vol. 51 No. 4, pp. 1029-1039, doi: 10.1177/001316449105100423.

Yap, K., Bearman, M., Thomas, N. and Hay, M. (2012), "Clinical psychology students' experiences of a pilot objective structured clinical examination", *Australian Psychologist*, Vol. 47 No. 3, pp. 165-173, doi: 10.1111/j.1742-9544.2012.00078.x.

Yousuf, N., Violato, C. and Zuberi, R.W. (2015), "Standard setting methods for pass/fail decisions on high-stakes objective structured clinical examinations: a validity study", *Teaching and Learning in Medicine*, Vol. 27 No. 3, pp. 280-291, doi: 10.1080/10401334.2015.1044749.

Further reading

Crocker, L. and Algina, J. (1986), *Introduction to Classical and Modern Test Theory*, Harcourt, New York, NY.

Downing, S.M., Tekian, A. and Yudkowsky, R. (2006), "Research methodology: procedures for establishing defensible absolute passing scores on performance examinations in health professions education", *Teaching and Learning in Medicine*, Vol. 18 No. 1, pp. 50-57, doi: 10.1207/s15328015tlm1801_11.

Kane, M.T. (1992), "An argument-based approach to validity", *Psychological Bulletin*, Vol. 112 No. 3, pp. 527-535, doi: 10.1037/0033-2909.112.3.527.

Messick, S. (1995), "Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, Vol. 50 No. 9, pp. 741-749, doi: 10.1037/0003-066X.50.9.741.

Tavakol, M. and Dennick, R. (2011), "Making sense of Cronbach's alpha", *International Journal of Medical Education*, Vol. 2, pp. 53-55.

Corresponding author

Anna E. Sundström can be contacted at: anna.e.sundstrom@umu.se