Assessing competencies of trainee sport psychologists: An examination of the ‘structured case presentation’ assessment method


MOVE Research Institute Amsterdam, Faculty of Behavioural and Movement Sciences, VU University Amsterdam, The Netherlands

ABSTRACT

Objectives: There is virtually no literature on how to assess competencies of applied sport psychologists. We assessed casework of applied sport psychology students and compared written case report assessment (WCRA) with structured case presentation assessment (SCPA) on reliability and acceptability (e.g., validity, transparency, feedback function and preference of methods, as perceived by students and assessors).

Design: A quantitative, comparative study of two assessment methods.

Method: Participants were 11 students, nine supervisors and three exam committee members. A number of 18 cases were evaluated with both WCRA by the supervisor and SCPA by two exam committee members. Ten of these cases were also evaluated with WCRA by exam committee members. Interrater reliability measures were calculated and compared for the different assessment methods. Participants' perception of the validity, transparency, and feedback function of the methods, and the preferences for assessment methods were surveyed with a brief questionnaire.

Results: SCPA by the exam committee resulted in higher interrater reliability than WCRA by supervisor and exam committee. The feedback function of SCPA seemed superior to WCRA by either supervisor or exam committee. For assessment by the exam committee, the perceived validity and transparency of SCPA seemed higher than of WCRA. Students and exam committee had the highest preference for SCPA by supervisor and exam committee.

Conclusions: Overall it can be concluded that, for assessment by the exam committee, structured case presentations provided a more reliable and acceptable method of assessment than written case reports only.

As the world of sport is becoming increasingly professional, it is important for the field of sport psychology to develop in tandem with this professionalization. The training and professional development of sport psychology practitioners play an important role in this process. In recent years, education and training of sport psychologists has received increasing attention. Insight into what should be learned to be successful in sport psychology has evolved through research, evaluation, and personal accounts (e.g., Cropley, Hanton, Miles, & Niven, 2010; Fifer, Henschen, Gould, & Ravizza, 2008; Gould, Murphy, Tammen, & May, 1991; Partington & Orlick, 1987; Simons & Andersen, 1995; Ward, Sandstedt, Cox, & Beck, 2005; Weigand, Richardson, & Weinberg, 1999; Yukelson, 2001). In these studies a wide variety of skills have been identified as important for sport psychologist's development, such as reflective practice, critical self-evaluation and abiding by ethical regulations, applying and understanding interventions, understanding and fitting in with the sporting context and culture, communication skills, and building an effective working alliance.

Following up on this body of literature, and in accordance with developments in related fields such as medicine and professional psychology, competencies for sport psychology service delivery have been conceptualised (AASP, 2012; APA, 2005; Fletcher & Maher, 2013; Stapleton, Hankes, Hays, & Parham, 2010; Tenenbaum, Lidor, Papaianou, & Samulski, 2003). Organisations, such as the American Psychology Association (APA, division 47), the Association for Applied Sport Psychology (AASP), and the International Society of Sport Psychology (ISSP), have outlined areas of competence, mostly distinguishing between required knowledge
and skills. ISSP, for example, depicted intervention skills and communication skills (Tenenbaum et al., 2003), and AASP (2012) distinguished between counseling skills and skills in sport and exercise, such as coaching. As for the knowledge-oriented competencies, knowledge of psychology, sport science, and sport psychology theory and practice were commonly listed competencies (AASP, 2012; APA, 2005; Tenenbaum et al., 2003). In addition, both ISSP and AASP explicitly mentioned knowledge of research and scientific methodologies.

The advanced knowledge of competencies is important to enhance training and performance in the professional field of sport psychology. However, one important aspect of competencies is still absent: There is virtually no literature on how to assess sport psychologists’ competencies, or how to evaluate whether trainee sport psychologists are “fit for practice”. For certification by AASP (2012), for example, a mentorship evaluation and verification form has to be completed by mentors, but instruction is lacking on how scores on the criteria on this form should be established. APA, division 47, provides a self-assessment checklist “intended to assist practitioners in assessing their knowledge and skills in terms of the proficiency criteria” (2005, p. 1), but no benchmarks or recommendations for self-assessment are given. ISSP (Tenenbaum et al., 2003) specified competencies in different competence standards, but also overlooked the matter of how these standards should be evaluated. In short, although competence domains and criteria have been described, the question of how to assess competence remains largely unattended.

One possible explanation for the dearth of literature on competency assessment in sport psychology could be that there are hardly any competency problems in the field, and the unspecified systems are actually working well. However, in 2000 Andersen, van Raalte, and Brewer stated “Given the research in clinical and counseling psychology and the limited studies among sport psychology consultants, impairment is an area waiting for research in the field of sport psychology” (p. 137). It seems that the wait is still on-going; a literature search on professional impairment, competency problems or gatekeeping in sport psychology rendered no studies investigating the occurrence and incidence of competency problems in sport psychology. Thus, it is evident that the process of developing and assessing professional competence has been largely overlooked within the sport psychology domain.

Preventing competency problems in the field may not be the only function of competency assessment, though. According to Kaslow (2004) “the assessment of competence fosters learning, evaluates progress, assists in determining the effectiveness of the curriculum and training program, and protects the public.” (p. 778). These are all functions that contribute to the quality of practice and education in sport psychology, and thus lead to the professional development in the field.

Competency assessment for complex professions is not an easy task. According to Epstein (2007) “All methods of assessment have strengths and intrinsic flaws” (p. 388). Similarly, Kaslow et al. (2009) highlighted “salient challenges” (p. 534) for all assessment methods in their competency assessment toolkit for professional psychology. In our personal experience, we recognize the difficulties in designing appropriate and acceptable measures for the evaluation of sport psychology students’ competence. We are, in different roles, responsible for assessment in the post-master program in applied sport psychology in the Netherlands. The first author is a member of the exam committee and teacher in the program. The second author is the program manager. The third author is a member of the steering committee of the program. The impetus for the current study was our wish to take a critical look at the assessment method applied in the program, and to investigate an alternative way of assessing competence. The study aims to contribute to the knowledge on methods to assess competence, in our case of trainee sport psychologists who are about to enter practice. As outlined previously, despite the important functions of assessment of competence, very little is known about competence assessment in the field of sport psychology.

1. Context and background of the study

The most important assessment of competence in our post-master program is the evaluation of supervised casework, with which the student trainees conclude the program. Before the study took place, cases were evaluated by supervisors and an exam committee by assessing a written report on the case (see also Hutter, 2014). Both students and assessors were critical of this written case report assessment (WCRA). Assessors had the impression that the written reports did not completely capture the how, what and why of the students’ professional actions (see also Hutter, 2014). Moreover, (wide) discrepancies occurred (in some cases) between the evaluation by the supervisor and the exam committee. Students indicated that they struggled to clearly present the professional decision-making and rationales behind their actions in the written reports. On a pragmatic level, both students and assessors perceived the written reports to be time consuming, tedious, and inefficient in terms of learning. Different actions were taken to improve the assessment. For example, we tried to reduce the discrepancies between assessors by revising the assessment criteria list and providing training for the assessors (Hutter, 2014). To try to relieve the workload associated with the written reports, we prescribed a maximum number of pages for the case reports. Although these actions improved some aspects of the assessment, room for improvement remained. Therefore we continued to discuss the needs, challenges, and available methods for assessment with stakeholders (such as students, assessors, and supervisors) and turned to the literature on competency assessment in search of a better assessment method. Petti (2008) offered an elaborate description of a structured case presentation procedure that is used at the California School of Professional Psychology to assess students’ clinical competence (from here on called CSPPA). For CSPPA, students hand in a written case report, which is read and (preliminarily) evaluated by the assessors. Next, the assessors conduct a 60 min oral examination with the student. At the end of the oral examination, the assessors complete the evaluation and grading of the student, using a standardised criteria list outlining different competencies and competency domains. The CSPPA (Petti, 2008; Swope, 1987) has been reported to have high reliability (Dienst & Armstrong, 1988; Tori, 1989, both as cited in Petti, 2008), and it has been used for over a decade (Petti, 2008). Goldberg, Delamatre, and Young (2011) compared CSPPA with three other assessment methods for the evaluation of performance of clinical psychology interns. CSPPA was the preferred method of both the interns and the examiners in the study, in terms of clarity, simplicity, and fidelity. Kaslow et al. (2009) categorised CSPPA as case review presentations, and discuss this method in the competency assessment toolkit for professional psychology. They deemed case review presentations useful for assessing different foundational competencies (e.g., interventions) at different developmental levels, including entry level to practice.

For our study, we adapted the CSPPA structure of assessment into what we labelled a Structured Case Presentation Assessment (SCPA; see ‘Method’ section for further explanation). We chose the CSPPA structure with an emphasis of the before mentioned stakeholders, for its evidence-based merit, but also its practical feasibility within the post-master program. Standardised criteria lists were already part of the assessment method applied prior to the study. For SCPA,
the same criteria lists could be used, with the criteria now scored in two stages; first, preliminarily on the basis of the written report and second, after the discussion meeting with the student. Organisation-wise, only the discussion meeting had to be added on top of the written reports used in the original WCRA method. The rest of the logistics of assessments did not require any substantial changes for SCPA to be introduced.

In general, in applying assessment methods, assessors should strive for high validity, reliability, objectivity, and transparency (e.g., van Berkel & Bax, 2015), for the provision of valuable feedback for the professional development of students (e.g., Hattie & Timperley, 2007), and for assessment methods that are highly accepted by the people involved (e.g., van der Vleuten, 1996). In the current study we compared evaluation of the written case reports only (i.e., WCRA) to evaluation on the basis of a written case report and a discussion meeting with the student (i.e., SCPA). Our comparison of assessment methods focused on interrater reliability and the preferences of students and supervisors.

2. Method

2.1. Participants

2.1.1. Students
At the start of the study, all students in the final stage of the post-master program (17 in total) received a written explanation of the study by e-mail, outlining the aim of the study and the rationale and procedure of the SCPA. The final stage of the program consists of supervised casework of at least seven different cases, and a minimum of 70 client contact hours. Each case consists of at least six client contact hours, and five hours of indirect supervision (i.e., the supervisor is not present at the client contact, supervision sessions are held with supervisor and student). During the research period, 11 students (5 women, 6 men, average age 27.2 years, SD = 1.8) concluded at least one of their first two cases. A total of 19 cases from these 11 different students were evaluated with both WCRAs and SCPAs. The first SCPA served as a pilot assessment, to allow the assessors to become familiar with the procedure, time-keeping, structure, and execution of SCPA. The remaining 18 SCPAs and corresponding WCRAs were included in the study.

2.1.2. Assessors
The 18 cases included in the study were supervised by nine different supervisors. The exam committee involved in the study consisted of three members; two members were also teachers in the post-master program, the third was also a supervisor in the program. All assessors were senior sport psychologists with much experience in sport psychology practice. They also had experience with the WCRAs, and had received basic training in WCRA. The assessors who conducted the SCPA, (i.e., the exam committee, see ‘SCPA procedure’ below) received information in writing and in person about the aim and procedure of the SCPA.

2.2. Inclusion of cases
The assessment of the first two cases of students in the program was formative, meaning that students obtained an evaluation to benchmark their level of competence. An insufficient grade (i.e., unsatisfactory) in this stage had no consequences for students, other than informing them that their level of competence was under par. The assessment of the remaining five cases was summative: Students had to obtain a grade of at least satisfactory to complete the particular case. For the current study, only formative assessed cases (i.e., the first two cases of students) were included. In this way, if different outcomes of both assessment methods would occur, this would not have undesired positive or negative practical consequences for students.

Prior to the study, all formatively assessed cases (i.e., first and second cases) were evaluated using WCRA by the supervisor, and each first case of a student was evaluated using WCRA by a member of the exam committee. During the study, all included cases were additionally evaluated by SCPAs. SCPAs were conducted by two members of the exam committee (in alternating pairs). The following evaluations were thus obtained for first cases of students: WCRA by the supervisor, WCRA by a member of the exam committee, SCPA by the same member of the exam committee, and SCPA by an additional exam committee member. For second cases of students, WCRA evaluation by the supervisor, and SCPA evaluations of two exam committee members were obtained. An overview of the included cases with corresponding assessors and assessments is presented in Table 1.

All cases that were concluded in the research period were included in the study, except one. For the excluded case, no SCPA meeting took place, due to logistic problems in planning the examination.

2.3. Assessment procedures

After completion of each case, students handed in a written report. The written report consisted of two parts: an overall case report (maximum length 12 pages), and reports of each session (maximum length 1 page per session).

The written reports were evaluated with two criteria lists, that is, the criteria list for case reports and the criteria list for session reports. The criteria lists (see Appendix) outline different competencies and domains and have been developed through the collaboration of exam committee members and supervisors (see Hutter, 2014). The criteria list for case reports consists of 39 items and was designed to evaluate six domains: intake session and presenting concern (10 items), professional vision and working model framework (seven items), goals (four items), treatment plan (five items), evaluation (four items), and reflective practice (nine items). The criteria list for session reports consists of 24 items and addresses three domains: session-goal (five items), interventions (nine items), and reflective practice and communication (10 items). Each item on the lists was scored with unsatisfactory, unsatisfactory/satisfactory, satisfactory, satisfactory/good, or good.

2.3.1. WCRA procedure
Written reports were handed in to the program manager and sent for evaluation to the assessors involved in the case (i.e., the supervisor, and if the report concerned the first case of a student, a member of the exam committee). The assessors completed the evaluation with the criteria lists for case and session reports and sent these back to the program manager. This was the customary procedure of assessment of cases in the post-master program, prior to the study.

2.3.2. SCPA procedure
For SCPA, the written reports were sent to the assessors (i.e., two members of the exam committee, for first cases one of them was the same member that completed the WCRA of the case). In line with the CSPPA (Petti, 2008), assessors were allowed ample time to read the report. In most cases the SCPA meeting took place within four weeks after handing in the report. Assessors were requested to evaluate the case report, and to evaluate one session report. The session to be evaluated was randomly selected by the program manager. The assessors scored the same criteria lists for case and session reports as in WCRA, but did so in two stages in the SCPA procedure. After reading the written case report, the assessors gave
scores for all items for which they felt they had sufficient information from the written report. If assessors felt that they lacked information to give a score for a particular item on the basis of the written report only, they marked the item as to be determined at the SCPA meeting. The assessors then completed their evaluations after the SCPA meeting took place. For the purpose of the study, the assessors sent both preliminary and final SCPA evaluations to the program manager.

For the majority of the evaluations the WCRA and SCPA were separate assessments. However, for first cases of students, one member of the exam committee evaluated the case twice. This member completed a WCRA based on the written report only; and a preliminary SCPA evaluation based on the written case report only, followed by a final SCPA based on the case report plus the SCPA meeting. Although the WCRA and SCPA are presented as separate procedures, in these cases of “double assessment” the WCRA scoring and preliminary SCPA scoring were, in reality, done in a combined fashion. The assessor typically first scored the written report with WCRA, with forced scores on items that they ideally would like or need more information on; and marking these items as to be determined at the SCPA meeting for the preliminary SCPA evaluation.

The structure of the SCPA meeting was the same as in the CSPPA (Petti, 2008): In the first 10 min the assessors prepared the meeting in absence of the student, then a 40 min period was used to discuss the case with the student, and the last 10 min were used to prepare and provide feedback to the student. Just as with CSPPA, the student did not have to present the case, because information was already available in the written report. The assessors used the preparation time to compare which scores they had indicated to be determined in SCPA meeting and sketched a rough guideline for the discussion time with the student. In the 40 min discussion with the student, the assessors tried to obtain information of the student’s competence on the criteria they felt they could not score on the basis of the written report only. Attempts were made to lower the students’ anxiety for the assessment by explaining the procedure thoroughly, by allowing them ample time to think and collect their thoughts and feelings, and by verbally and nonverbally encouraging open and honest reflections about the case. Moreover, in each SCPA meeting the assessors clarified that the points that would be discussed were on the table because the assessors needed more information to determine the level of the students’ competence, and not necessarily because these were weak points in the case.

After the discussion with the student, the assessors briefly discussed the meeting separately from the student, and prepared key points of feedback. Next, the assessors delivered this feedback verbally to the student. Finally, the student was invited to share his/her experience of the SCPA meeting with the assessors. After the SCPA meeting (usually the same day) the assessors completed the criteria lists for case and session reports (SCPA scores), scoring the items that were previously marked as to be determined in SCPA meeting as well, and sent these to the program manager. The program manager then sent all completed criteria lists of the case to the student. In exceptional cases in the study, the student had received the evaluation from the WCRAs before the SCPA meeting took place.1 After assessments were completed, participants received an evaluation questionnaire and gave informed consent for their data to be included in the study.

2.4. Data collection

2.4.1. Scores of criteria lists

The following data were collected for all cases ($N = 18$): (a) WCRA scores of the supervisor for all items on the criteria lists; (b) the to be determined in SCPA meeting scores on the criteria lists of two exam committee members; and (c) SCPA scores of two exam committee members for all items on the criteria lists. In addition, the WCRA scores of an exam committee member were collected for the students’ first cases included in the study ($n = 10$).

2.4.2. Opinion of students and assessors on assessments

Students and assessors were requested to complete a brief online questionnaire. The first author (as a member of the exam committee) did not complete the questionnaires, to prevent researcher bias in these data. Respondents were asked to indicate on a visual analogue scale the extent to which they agreed with statements addressing:

\[\text{Education and examination regulations state that students are entitled to obtain evaluations within four weeks after completion. Therefore, if a SCPA could not be scheduled within four weeks after handing in a report, and the student claimed his/her right to obtain the evaluations, then the completed WCRA were sent to the student, even though the SCPA meeting or the SCPA evaluation had not been completed.}\]
• transparency (the evaluation was completed in a fair fashion [students]/I was able to evaluate the case objectively [assessors], and the grounds on which I am evaluated are clear to me [students]/the grounds on which I should evaluate the case are clear to me [assessors]);
• validity (the evaluation does justice to my professional competence during this case [students]/I was able to judge the level of professional competence of the student during this case [assessors], and I agree with the evaluation [students]/I stand by my evaluation [assessors]);
• feedback (the feedback I received is helpful for my professional development [students]/the feedback I provided is helpful for the students’ professional development [assessors], and I can relate to the feedback [students]/I stand by my feedback [assessors]).

The online visual analogue scales rendered scores ranging from zero (no agreement with the statement) to one hundred (full agreement with the statement).

Students scored the six statements for each assessment method with which their case was evaluated (i.e., WCRA by the supervisor, WCRA by the exam committee [after first cases], and SCPA by the exam committee). Supervisors scored the six statements for their WCRA evaluations. Exam committee members scored the statements for the assessments they applied to a case (i.e., WCRA and/or SCPA).

In the last question of the questionnaire, students and exam committee members were asked to rank different assessment methods in order of preference. We wanted to check which assessment method (WCRA or SCPA) and which assessors (supervisors, exam committee members, or both) were preferred by students and assessors. These two dimensions (assessors and method) combined, resulted in six options to be ranked:

- WCRA by supervisor only
- WCRA by exam committee only
- WCRA by supervisor and exam committee
- SCPA by supervisor only
- SCPA by exam committee only
- SCPA by supervisor and exam committee

2.5. Data analysis

2.5.1. Frequencies of the score to be determined in the SCPA meeting

Frequencies of the score to be determined in the SCPA meeting were calculated for each item on the criteria lists. Exam committee members gave this score in preparation of SCPA, when they felt that the written report contained too little information to be able to score the item. In the WCRA, assessors are “forced” to give a score for each item. Frequencies of to be determined in the SCPA meeting thus indicate how often judgements are forced to be made with insufficient information. We argue that scores that are given on the basis of insufficient information are unreliable. Therefore, we used the frequencies of the scores to be determined in the SCPA meeting as indicators of reliability of WCRA.

2.5.2. Interrater reliability

Krippendorff’s alpha (K-alpha) was used as a measure of interrater reliability. K-alpha is frequently used in qualitative research and incorporates characteristics of other reliability/agreement measures such as Fleiss’ kappa, Scott’s pi, or correlation coefficients (Hayes & Krippendorff, 2007). K-alpha takes agreement by chance into account, is suitable for ordinal data, and can be applied to datasets with missing data (Krippendorff, 2011). K-alpha ranges from –1 to 1. If the judgements are unrelated (i.e., agreement achieved is equal to achievement that can be expected by chance) K-alpha is 0. A negative K-alpha indicates that disagreements are systematic: The agreement is worse than what would result from chance. Krippendorff (2004) contends that a K-alpha of at least .80 indicates perfect interrater agreement, but tentative conclusions are acceptable for K-alphas of at least .667. He states that “Except for perfect agreement, there are no magical numbers, however.” (p. 429).

2.5.3. Types of agreement

Interrater reliability was calculated for three types of agreement:

- WCRA-agreement, which is a comparison of the WCRA scores of the supervisor with the WCRA scores of the member of the exam committee (students’ first cases only, n = 10, a total of 630 pairs);
- SCPA-agreement, which is a comparison of the SCPA scores of the two exam committee members (N = 18, a total of 1134 pairs);
- WCRA/SCPA-agreement, which is a comparison of the WCRA scores of the supervisor with the SCPA scores of the two exam committee members (students’ first cases only, n = 10, a total of 3402 pairs).

2.5.4. Levels of calculation

For each type of agreement we calculated interrater reliability on two levels: overall, and separately per item for the 63 items on the criteria lists. With the overall calculation the general interrater reliability of each assessment method is evaluated. With the calculation per item interrater reliability of separate items are inspected. It could be, for instance, that certain items are weak in terms of interrater reliability, while other items render high agreement between assessors. Interrater reliability was calculated on both levels, for each type of agreement, resulting in the following K-alpha’s: K-alpha WCRA-agreement overall, K-alpha SCPA-agreement overall, 63 K-alphas for WCRA-agreement (one for each item), 63 K-alphas for SCPA-agreement, and 63 K-alphas for WCRA/SCPA-agreement.

2.5.5. Statistical comparison of interrater reliability between types of agreement

To analyse whether interrater reliability differed between SCPA and WCRA, we compared the K-alphas for SCPA agreement per item with the K-alphas for WCRA agreement per item, using Wilcoxon signed rank tests for paired samples.

Moreover, we wanted to check whether SCPA by the exam committee resulted in better agreement with the supervisor’s (WCRA) evaluation of the case than WCRA by the exam committee. After SCPA, exam committee members have more information on the case and the student. Therefore their evaluation might be more equal to the evaluation of the supervisor. We tested whether there was a difference between K-alphas for WCRA/SCPA agreement per item and for WCRA agreement per item, using Wilcoxon signed rank tests for paired samples.

2.5.6. Questionnaire data

Finally, we examined the questionnaire data of students and assessors to evaluate the validity, transparency, and feedback function of each assessment method as perceived by assessors and students, and their preference for assessment methods. For each assessment method in the study (WCRA by supervisor, WCRA by exam committee, and SCPA by exam committee) descriptive statistics were calculated for each statement in the questionnaire. For
the ranking of assessment methods, frequencies of ranks were calculated, that is, how often an assessment method was ranked as first preference, second preference, etc.

3. Results

3.1. Frequencies of the score ‘to be determined in the SCPA meeting’

A total of 36 evaluations were completed by members of the exam committee (18 cases, assessed by two members). Frequencies of the score to be determined in the SCPA meeting were calculated for each item on the criteria lists. These frequencies ranged from 0 to 24 (out of the 36 evaluations). On average, an item required further discussion at the SCPA meeting in 32.5% of the evaluations. However, there are substantial differences among items in the number of times they had to be discussed at the SCPA meeting to determine a score (SD = 16.4). For some items, exam committee members were generally able to give a score based on the written report. The following items were almost always (>94% of the evaluations) given a score without requiring discussion at the SCPA meeting: describes the sports history of the client, establishes and describes a case-concept of the client, and describes the psycho-diagnostic assessment in line with the standards of the course psycho-diagnostics.

A number of items (n = 9) were frequently (>55% of the evaluations) scored as to be determined in SCPA meeting, thus indicating that the exam committee often struggled to give a score on the basis of the written report only. These items are: notices potential discrepancies within the behaviour of the client, and acts upon it; adapts consulting style and approach to the client; varies consulting style and approach with client; substantiates the choice for the consulting style and approach; recognizes potential transference and countertransference processes; establishes whether the skill and practice material is understood by the client; describes possibilities for improvements related to the strong and weak points; is guided by a clear working vision in the session; and applies issues that were addressed in supervision in the session.

On average, exam committee members scored 21.7 (SD = 6.0) out of the 63 items on the criteria lists as to be determined in SCPA meeting. This indicates that the written report provided the exam committee with too little information to give an appropriate score for about one third of their scores. The exam committee members did not always agree whether an item should be discussed at the SCPA meeting to determine a score. There were 1134 pairs of scores in total: 18 cases, scored on 63 items. In about half of these pairs, exam committee members agreed that a score could be given without further discussion at the SCPA meeting (565 pairs, 49.8%). In about one sixth of the pairs, the exam committee members agreed that discussion at a SCPA meeting was required to determine a score (155 pairs, 13.7%). However, in 414 instances (36.5%) one exam committee member indicated that discussion at a SCPA meeting was required to determine a score, while the other member gave a score without need for further discussion.

Overall the results show that, according to the exam committee, the written report is often insufficient to give an appropriate score for items on the criteria lists. This finding raises serious concerns about the reliability of evaluation of cases based on the written report only.

3.2. Interrater reliability

Interrater reliability was highest for SCPA and lowest for WCRA. The K-alphas overall were .42, .25, and .20 for SCPA-agreement, WCRA/SCPA agreement, and WCRA agreement, respectively. The lowest K-alpha for WCRA agreement further adds to the concern on reliability of WCRA. K-alphas of all assessment methods were below .667, the level required for tentative conclusions, but the agreement between scores was higher for SCPA evaluations than for WCRA evaluations. The agreement between SCPA evaluations of the exam committee and the WCRA of the supervisor was higher than the agreement between exam committee and supervisor with WCRA, but lower than the SCPA agreement between exam committee members.

K-alphas were calculated per item on the criteria lists, and compared for SCPA and WCRA. See Fig. 1 for a summary of the K-alphas per item. There was a significant difference between K-alphas per item for SCPA agreement (Mdn = .43), and the K-alphas per item for WCRA agreement (Mdn = .17; Z = 4.57, p < .001). Interrater reliability of SCPA by exam committee members was superior to the interrater reliability of WCRA by supervisors and exam committee members.

Next, we checked whether the scores of exam committee members were in better agreement with the supervisors’ scores with SCPA than after WCRA. We compared the K-alphas per item of WCRA/SCPA agreement to the K-alphas per item of WCRA agreement. See Fig. 2 for a summary of the K-alphas per item. There was no significant difference between WCRA/SCPA-agreement (Mdn = .18) and WCRA-agreement (Mdn = .17; Z = 1.24, p = .212). After SCPA the scores of the exam committee were in better agreement with the supervisor scores than after WCRA.

There were a number of outliers in the K-alphas per item. These outliers indicate items that are scored with exceptionally high, or exceptionally low, interrater agreement. In WCRA interrater reliability was low for the following items: reflects on the consequences of these thoughts and feelings for his/her professional actions; describes strong and weak points of the session; and notices potential discrepancies between the behaviour of the client and the case-concept, and acts upon it. In SCPA, agreement was low for the items: integrates the conclusions from the psycho-diagnostic assessment into the treatment/intervention outline, and recognizes potential transference and countertransference processes (session). Relatively high agreement existed between SCPA scores of the exam committee and WCRA scores of supervisors (WCRA/SCPA agreement) for the item formulates a clear and realistic intervention plan.

3.3. Questionnaire data

After each assessment, students and assessors gave their
Students and exam committee members ranked six different combinations of assessment methods and assessors in order of their preference. The ranks given by students and exam committee are presented in Fig. 3. As can be seen in the figure SCPA received more often high ranks (first to third preference) than WCRA, which received more low ranks (fourth to sixth preference). The results thus illustrated a clear preference of both students and exam committee members for SCPA over WCRA. Moreover, both students and exam committee members valued a combination of the supervisor and an exam committee member as assessors. The most preferred assessment method is SCPA by the supervisor and a member of the exam committee, followed by SCPA by exam committee members, as applied in the study.

4. Discussion and conclusions

In the current study, we investigated whether SCPA is a better assessment method than WCRA. As expected, we found that reliability of WCRA was problematic. We found that reliability of assessment was better with SCPA, and that, for assessment by the exam committee, students and assessors perceive SCPA to have benefits in terms of validity, transparency and feedback function. We will discuss these findings in more detail, starting with the comparison of interrater reliability and the questionnaire data between the assessment methods, followed by a discussion of the limitations and implications of the study.

Kaslow et al. (2009) stated that it is challenging to assess skills and attitudes and to achieve high fidelity with written examinations. Moreover, they pointed out that record reviews (comparable to the session reports assessed in the study) may fail to capture all procedures, interventions, and treatment components. They also stressed that direct observation data are required for reliable assessment with competency evaluation rating forms. Our findings are in line with Kaslow et al.’s critique of written examinations. We found that often information was lacking in WCRA for proper assessment, and that interrater reliability of WCRA was low.

Interrater reliability was significantly higher with SCPA than with WCRA. In SCPA, assessors hear the students describe knowledge application, skills, and values in interaction with clients, and can observe verbal and nonverbal communication (Kaslow et al., 2009). Apart from the extent to which the evaluation does justice to the level of competence, students valued SCPA by exam committee equally, or higher than WCRA by their supervisor.

Students and exam committee members ranked six different combinations of assessment methods and assessors in order of their preference. The ranks given by students and exam committee are presented in Fig. 3. As can be seen in the figure SCPA received more often high ranks (first to third preference) than WCRA, which received more low ranks (fourth to sixth preference). The results thus illustrated a clear preference of both students and exam committee members for SCPA over WCRA. Moreover, both students and exam committee members valued a combination of the supervisor and an exam committee member as assessors. The most preferred assessment method is SCPA by the supervisor and a member of the exam committee, followed by SCPA by exam committee members, as applied in the study.

Table 2
Mean scores and standard deviations (between brackets) of the extent to which students and assessors agree with the statements about transparency, validity, and feedback function of each assessment method.

<table>
<thead>
<tr>
<th>Students</th>
<th>Respondents</th>
<th>N of responses</th>
<th>11</th>
<th>9</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment method</td>
<td>WCRA</td>
<td>WCRA</td>
<td>SCPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>By Sup</td>
<td>By EC</td>
<td>By EC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The evaluation was completed in a fair fashion</td>
<td>73.8 (22.8)</td>
<td>62.1 (25.1)</td>
<td>72.3 (26.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The grounds on which I am evaluated are clear to me</td>
<td>66.5 (28.9)</td>
<td>65.3 (24.6)</td>
<td>73.2 (29.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The evaluation does justice to my prof. competence during this case</td>
<td>71.4 (21.5)</td>
<td>52.0 (33.0)</td>
<td>66.3 (32.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I agree with the evaluation</td>
<td>74.3 (17.8)</td>
<td>64.2 (27.3)</td>
<td>73.4 (28.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The feedback is helpful for my prof. development</td>
<td>56.8 (36.5)</td>
<td>63.8 (28.3)</td>
<td>79.0 (24.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can relate to the feedback</td>
<td>62.3 (32.8)</td>
<td>61.9 (30.0)</td>
<td>74.5 (28.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessors</td>
<td>Respondents</td>
<td>N of responses</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Assessment method</td>
<td>WCRA</td>
<td>WCRA</td>
<td>SCPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>By Sup</td>
<td>By EC</td>
<td>By EC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to evaluate the case objectively</td>
<td>67.7 (22.5)</td>
<td>81.4 (5.6)</td>
<td>89.4 (6.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The grounds on which I should evaluate the case are clear to me</td>
<td>73.7 (18.4)</td>
<td>78.4 (9.2)</td>
<td>88.1 (6.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to judge the level of prof. competence during this case</td>
<td>70.1 (27.6)</td>
<td>71.2 (9.8)</td>
<td>88.9 (8.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I stand by my evaluation</td>
<td>84.4 (15.9)</td>
<td>83.8 (9.6)</td>
<td>95.6 (8.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The feedback is helpful for the students’ prof. development</td>
<td>81.9 (17.2)</td>
<td>84.2 (9.2)</td>
<td>86.4 (12.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I stand by my feedback</td>
<td>86.8 (15.3)</td>
<td>89.2 (8.0)</td>
<td>92.6 (9.4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
All these factors may contribute to the higher interrater reliability of SCPA. Another factor of importance concerns the assessors involved with each method. For the interrater reliability of WCRA, the evaluation of the supervisor is compared with the evaluation of a member of the exam committee. These two assessors differ in the amount of contextual knowledge they have on the student and the case, in that one has been supervising the case, the other has only read the students’ written report. In SCPA the knowledge of the case and student is the same for both assessors, as both members of the exam committee have read the written report and participated in the SCPA meeting. This homogeneity of information may play a role in the higher interrater reliability of SCPA.

In the discussion on interrater reliability it is important to bear in mind that, although intuitively the assessment of the supervisor may have higher fidelity to students and assessors (and some of the data point in that direction), substantial direct observation of trainees’ interactions with clients is required for proper evaluation in field settings (Pulito, Donnelly, Plymale, & Mentzer, 2006, as cited in Epstein, 2007). In our set-up, where supervision is indirect, the supervisor has no direct observation of the interaction. Moreover, different studies have addressed the potential bias of supervisors in evaluation of supervisees (e.g., Gonzalez et al., 2013; Yap, Bearman, Thomas, & Hay, 2012). Thus, rather than privileging the supervisor’s assessment as the gold standard and expecting the exam committee members to assimilate their assessments, we strive for assessments from both assessors that are as objective as possible and result in high agreement between evaluations.

Although interrater reliability is higher for SCPA than for WCRA, the K-alfas are still lower than .667 recommended for tentative conclusions (Krippendorff, 2004). A low interrater reliability is a common problem in assessments. Jonsson and Svingby (2007) concluded that both consensus estimates (e.g., percent agreement or Cohen’s kappa) and consistency estimates (e.g., Spearman or Pearson correlations) often do not meet the minimum values required for acceptable judgements. Unfortunately, no interrater reliability of the original CSPPA is reported by Petti (2008). Therefore, the reliability found in the current study cannot be compared with previous applications of the same assessment procedure.

We conclude that the introduction of SCPA significantly improves interrater reliability of our assessment of casework. Reliability might be further improved through more elaborate training of assessors, exchanging and discussing evaluations among assessors, and operationalizing scoring criteria, for example in the form of rubrics.

We used the questionnaire data to compare the (perceived) validity, transparency, and feedback function of the assessment methods. Two different comparisons will be discussed: a comparison between different assessors (i.e., WCRA by the supervisor and WCRA by the exam committee) and between different methods (i.e., WCRA by the exam committee and SCPA by the exam committee).

WCRA by the supervisor is perceived to be more fair and valid than WCRA by the exam committee. The supervision relationship, information exchange during the case, and the guidance of the supervisor in the execution of the case, may all play a role in the better reception of the supervisor’s WCRA evaluation. Interestingly, the students seem to find the feedback that they receive from the exam committee (both with WCRA and SCPA) to be more helpful for their professional development than the feedback from the supervisor. According to Hattie and Timperley (2007): “Feedback is one of the most powerful influences on learning and achievement” (p. 81). They stated that feedback should address the three questions of where am I going, how am I going, and where to go next. The results suggest that the exam committee’s feedback was more informative for students regarding these three questions. It might be that the feedback of an external, third person is a valuable addition to the feedback already gained in the supervision context.

From the results of the questionnaire data it became clear that when students are assessed by the exam committee, they prefer the validity, transparency and the feedback function of SCPA over WCRA. The questionnaire items consistently received higher scores for SCPA than WCRA by the exam committee.

When asked to rank the assessment methods, both students and exam committee members expressed a clear preference for SCPA over WCRA. This finding is in line with Goldberg et al. (2011), who reported that SCPA was the preferred method of assessment of students and assessors in comparison to three other oral examination procedures. Students and exam committee members expressed the highest preference for SCPA to be conducted with the supervisor and an exam committee member. It seems that students (and exam committee members alike) prefer assessment by multiple assessors. This preference is contingent with recommendations in the literature to use multiple assessors (Baartman, 2008) or multisource feedback (Andrews, Violato, Al Ansari, Donnon, &
Pugliese, 2013) in assessment of competency. Based on the questionnaire data and the ranking, we conclude that SCPA provides an acceptable assessment method.

4.1. Limitations of the study

For practical, logistical and financial reasons we chose not to include supervisors in SCPA. Therefore, we cannot separate the influence of the assessment method (SCPA versus WCRA) from the influence of the assessors (exam committee versus supervisors) on the interrater reliability outcomes.

We concluded that WCRA by the exam committee is problematic, because often the written report left them with too little information to score the criteria. In the current study we did not investigate whether the same would be true for WCRA by supervisors. Since supervisors only completed WCRA, and not SCPA, we have no data on how often supervisors felt they could not give a score based on the information they had, and whether proper scoring would ideally require additional discussion. Without this information it is unclear whether the combination of the report with the information gained in supervision suffices for proper WCRA by supervisors. It would be interesting to investigate whether a SCPA meeting, centred around assessment instead of supervision, would be as beneficial for assessment by supervisors as it seems to be for assessment by the exam committee.

Another limitation is that we chose to survey the opinion of students and assessors after each completed assessment, instead of once for each student or assessor in the study. We envisaged that the perception of validity, transparency, and feedback function of the assessments might differ from assessment to assessment, or might change with experience. As a downside, however, fair statistical comparison is complicated with multiple data from single participants. The lack of statistical comparison means that only cautious conclusions regarding validity, transparency, and feedback function could be drawn from the study.

5. Implications

5.1. Implications for the post-master program

The results of the study have led to changes in the assessment of casework in the post master program. The program management values assessment by different assessors (i.e., both supervisor and exam committee) and this preference is shared by students and assessors. The main question therefore is how to retain the added value of multiple assessors, while overcoming the obstacle that exam committee members have relatively little information available for their evaluation. SCPA seems to offer a potential solution, since assessors after each completed assessment, instead of once for each student or assessor in the study. We envisaged that the perception of validity, transparency, and feedback function of the assessments might differ from assessment to assessment, or might change with experience. As a downside, however, fair statistical comparison is complicated with multiple data from single participants. The lack of statistical comparison means that only cautious conclusions regarding validity, transparency, and feedback function could be drawn from the study.

As outlined in the introduction, there is currently a dearth of literature on the assessment of competencies in sport psychology. Our study illustrates that critical evaluation of assessment methods and comparison between methods can lead to improvement of assessment in sport psychology education. More studies, preferably on a variety of assessment methods, and for different levels of competence, are needed to further advance assessment, education, and qualification in our field. To shape such studies, we may make use of the evidence available in related fields, such as medicine and professional psychology. Recommendations from this field include the use of multisource feedback (e.g., Andrews et al., 2013), assessment programs rather than single methods (e.g., Dijkstra, van der Vleuten, & Schuwirth, 2009; Epstein, 2007; Schuwirth & van der Vleuten, 2011), descriptions of trainee performance levels in vignettes (Gonsalvez et al., 2013), and standardised situations such as standardised role-play (Muse & McManus, 2013), computer simulation (M. L. Newell, Newell, & Looser, 2013), or objective structured observations (Yap et al., 2012). We took the CSSPA (Petti, 2008) from the field of clinical psychology, and tested this method in our sport psychology program, which led to the conclusion that the method is applicable and useful in sport psychology.

Apart from transferring specific methods from related fields, we may look at the process through which assessment of competence has progressed in these related fields. Of particular interest is the culture shift that has taken place in professional psychology (Roberts, Borden, Christiansen, & Lopez, 2005). In short, the route taken in professional psychology is that competencies were clearly established (e.g., Rodolfa et al., 2005), operationally defined (e.g., Foud et al., 2009), and appropriate methods to measure the competencies established (e.g., Kaslow et al., 2009), Fletcher and Maher (2013) recommended following a similar path in sport psychology, that is, to progress to clearer consensus and better operationalization of competencies, and to discuss and investigate assessment methods. The current study aimed to contribute to this recommended journey in sport psychology in a number of ways. First, in the study our assessment criteria were shared, which may contribute to the debate on, and operationalization of competency domains, and competency components. Second, in the study different methods of assessment were investigated (i.e., written case report assessment and structured case presentations), contributing to the knowledge base of different assessment methods. The study illustrates that assessment of competency can be improved when written examinations are substituted by structured case presentations. SCPA seems particularly useful for assessment by external assessors, for example exam committee members, or other non-supervising assessors. Third, our study and
measure for coding data. Communication Methods and Measures, 1, 77–89.


