The evaluation of a frame-of-reference training programme for intern psychometrists

Authors:
Gerdi Mulder[1]
Lené I. Jorgensen[1]
J. Alewyn Nel[1]
Deon Meiring[2]

Affiliations:
[1]WorkWell: Research Unit for Economic and Management Sciences, North-West University, Potchefstroom Campus, South Africa

[2]Department of Human Resource Management, University of Pretoria, South Africa

Correspondence to:
Alewyn Nel

Email:
alewyn.nel@nwu.ac.za

Postal address:
PO Box 114, School of Human Resource Sciences, North-West University, Potchefstroom, South Africa, 2520

Dates:
Received: 30 Oct. 2012
Accepted: 03 June 2013
Published: 30 July 2013

How to cite this article:
Mulder, G., Jorgensen, L.I., Nel, J.A., & Meiring, D. (2013). The evaluation of a frame-of-reference training programme for intern psychometrists. *SA Journal of Human Resource Management/SA Tydskrif vir Menslikehulpbronbestuur, 11*(1), Art. #506, 10 pages. http://dx.doi.org/10.4102/sajhrm.v11i1.506

Orientation: The use of assessment centres (ACs) has drastically increased over the past decade. However, ACs are constantly confronted with the lack of construct validity. One aspect of ACs that could improve the construct validity significantly is that of assessor training. Unfortunately untrained or poorly trained assessors are often used in AC processes.

Research purpose: The purpose of this research was to evaluate a frame-of-reference (FOR) programme to train intern psychometrists as assessors at an assessment centre.

Motivation of study: The role of an assessor is important in an AC; therefore it is vital for an assessor to be able to evaluate and observe candidates' behaviour adequately. Commencing with this training in a graduate psychometrist programme gives the added benefit of sending skilled psychometrists to the workplace.

Research design, approach and method: A quantitative research approach was implemented, utilising a randomised pre-test-post-test comparison group design. Industrial Psychology postgraduate students (N = 22) at a South African university were used and divided into an experimental group (n = 11) and control group (n = 11). Three typical AC simulations were utilised as pre- and post-tests, and the ratings obtained from both groups were statistically analysed to determine the effect of the FOR training programme.

Main findings: The data indicated that there was a significant increase in the familiarity of the participants with the one-on-one simulation and the group discussion simulation.

Practical/managerial implications: Training intern psychometrists in a FOR programme could assist organisations in the appointment of more competent assessors.

Contribution/value-add: To design an assessor training programme using FOR training for intern psychometrists in the South African context, specifically by incorporating this programme into the training programme for Honours students at universities.

Introduction

The popular use of Assessment Centres (ACs) has over the years drastically increased at international level in various applied industries (International Task Force on Assessment Centre Guidelines, 2010; Krause & Gebert, 2003). It is widely accepted that ACs are mostly used in the field of personnel psychology for processes such as recruitment, selection and identification of managerial potential and talent (Dilchert & Ones, 2009; Lievens & Thornton, 2005). Lievens and Thornton (2005) emphasise the efficacy and importance of the implementation of ACs in personnel selection and promotion. Although for a long time ACs were solely used at international level, this technique began to be established in South Africa as a popular assessment technique in 1974 (Meiring, 2008). Major companies incorporated ACs as a means of assessment, which led to a need for practitioners to exchange ideas in a constructive manner, and hence the founding of the Assessment Centre Study Group (ACSG) (Meiring, 2008). Since 1970 the main aim of the ACSG has been to hold annual conferences to promote new research, insights and the teaching of ACs in a constructive and effective manner.

Thornton and Rupp (2006) explain that an AC can be seen as a combination of work-like exercises as well as other assessment type procedures specifically designed to activate certain behaviour in candidates in order for those behaviours and skills to be evaluated and observed. Schlebusch (2008) claims that the main aim and purpose of an AC is to select the most appropriate participant to be appointed in a position or programme and also states that one of the criteria for an AC is that participants should be informed that results will influence the decision of appointment. Some specific features that should also be present in an AC are: a job analysis should be carried out; multiple simulations and assessment instruments should be utilised; multiple and competent observers and role-players should be present; behavioural and not psychological constructs should

be observed; behaviour should be noted and classified; data integration should take place and efficient feedback should be provided to participants (Schlebusch, 2008).

Although an AC is one of the more costly techniques used for assessment, ACs have good predictive validity (Eurich, Krause, Cigularov & Thornton, 2009; Thornton, Murphy, Everest & Hoffman, 2000) and criterion-related validity (Arthur, Day, McNelly & Edens, 2003). ACs also show evidence of good inter-rater reliability, although this also depends on the expertise level of the assessors (Lievens, 2002). Moreover, Joiner (2004) states that in the American private sector ACs reach a 300% return on investment (ROI) at some point.

Thornton and Mueller-Hanson (2004) state that although ACs consistently demonstrate criterion validity, the construct validity is still lacking in many instances. Collins, Schmidt, Sanchez-Ku, Thomas, McDaniel and Le (2003) mention in their study that evidence against construct validity, such as constant low construct validity in certain dimensions, has in fact been reported. The issue of construct validity can be seen as one of the biggest challenges that ACs face (Guion, 1998). In his study Lievens (2009, p.104) also mentions the significant issue of construct validity; he feels that ACs have to overcome the 'lack of evidence to measure the constructs (dimensions) they are reported to measure'. It can thus safely be said that the biggest unresolved problem that still remains in the practice of ACs is that of construct validity.

The consistency of assessor judgments is one specific aspect of ACs that influences or contributes to construct validity (Pell, Homer & Roberts, 2008). The main task of an assessor in an AC is to observe a candidate's behaviour and assign a rate, which then determines whether the candidate is appointed in a specific post (Goodstone & Lopez, 2001). The assessor's expertise therefore plays a significant role in the construct validity of the process (Jones & Born, 2008).

Assessors in assessment centres

The main task of ACs is the evaluation of various competencies, and a team of assessors is needed to observe and assess these competencies (Schlebusch, 2008). According to the International Task Force on Assessment Centre Guidelines (2010, p. 10), an assessee is 'an individual whose competencies are measured by an assessment centre'. This corresponds with previous research (Lievens, Tett & Schleicher, 2009; Schleicher, Day, Mayes & Riggio, 2002). Goodstone and Lopes (2001) confirm this by stating that an assessor's task is ultimately that of performance appraisal; therefore the essential part of any AC process is that of a trained assessor observing a candidate's behaviour and assessing it by giving it a rating.

The importance of validity in ACs is clear from findings by Jones and Born (2008) who found that assessors react more positively to behaviours and situations with which they are familiar – they therefore give emotive ratings. Schlebusch (2008) argues that up until now South African research has been reactive rather that proactive and that research on ACs, and specifically assessor training for the South African context, is limited. It is clear that although many issues contribute to the construct validity debate, one crucial element is that of assessors and their training.

Lievens (2009) asserts that trained observers should be used to observe participants in a typical job-related setting, whilst paying attention to various determined dimensions. Observing and evaluating participants are therefore carried out by observers or individuals otherwise known as assessors. Schlebusch (2008) defines assessors as the group of individuals who 'have the greatest impact on the whole assessment process'. Literature indicates that two of the most common mistakes made in any AC is, firstly, using unqualified assessors and, secondly, using poorly trained assessors (Thornton & Meuller-Hanson, 2004). Both Holmboe (2004) and Lievens (1998) found that the training of assessors could possibly have a significant effect on the construct validity of ACs. That the focus of assessor training should be on the quality of the training rather than the quantity (length) has been supported by research (Jackson, Atkins, Fetcher & Stillman, 2005). Schlebusch (2008) supports this statement when he states that not only the validity but also the reliability of an AC can be influenced by the quality of assessor training, and therefore specific care should be taken to ensure that they are indeed competent.

Training of Assessors

The main aim of training observers is to develop certain abilities that enable them to accurately and effectively rate participants' behaviour (Schlebusch, 2008). Lievens *et al.* (2009) stress the fact that sufficient training for assessors is critical. For these assessors to be able to rate accurately, Schlebusch (2008) says, some of the skills relevant to observing, noting, classifying and evaluating participants' behaviour during exercises or simulations have to be developed. They should also be able to record behaviour and reactions in detail and with precision. Schlebusch (2008) indicates steps that should ideally be followed for an individual who wishes to be classified as a competent assessor. Jones and Born (2008) claim that the level of assessor expertise significantly affect the validity of ACs and can be very beneficial to the AC process.

Schlebusch (2008) recommends that, in the South African context, an assessor in training should first attend an AC as a participant and then as an assessor (although their input will not be considered at that time). When individuals have attended two ACs, the International Task Force on Assessment Centre Guidelines (2010) advises they undergo lecture room training, after which they should twice act as assistant assessors under the supervision of a qualified and competent assessor (Schlebusch, 2008). The expert assessor, the AC administrator and other members of the assessor team all have to agree on the matter before the individual can be declared a competent assessor (International Task Force on Assessment Centre Guidelines, 2010; Schlebusch, 2008).

Lievens *et al.* (2009), however, claim that evidence exists for another technique, namely frame-of-reference training, which could increase inter-reliability, dimension differentiation and even criterion validity. Jackson *et al.* (2005) suggest that frame-of-reference (FOR) training should be implemented in the training of assessors to ensure a shared understanding of dimensions being measured.

Frame-of-reference training

Frame-of-reference (FOR) training focuses on developing a mutual understanding or frame of reference amongst assessors (Lievens, 2002; Lievens *et al.*, 2009; Schleicher *et al.*, 2002). The purpose of developing this mutual understanding is to equip all assessors with the same performance model that they can then utilise as a tool whilst observing at an AC (Lievens *et al.*, 2009). This mutual understanding can be reached by defining the dimensions (constructs or competencies) to be evaluated, providing and describing appropriate behavioural examples of the dimensions (constructs or competencies) to be evaluated, providing opportunities for practising evaluations, and finally providing feedback to assessors relating to their evaluations (Bernarding, Buckley, Tyler & Wiese, 2000; Melchers, Lienhardt, Von Aardburg & Kleinmann, 2011; Sulsky & Kline, 2007). The ultimate goal of FOR training is therefore to assist assessors in their tasks of observing and evaluating behaviours, and then categorising their observations into accurate and appropriate performance dimensions.

Lievens (2002; 2009) and Thornton and Rupp (2006) have on numerous occasions emphasised the importance and advantage of FOR training in increasing the effectiveness of assessors. Jackson *et al.* (2005) state that an explanation for this could be the fact that FOR training promotes an improved theoretical as well as practical understanding of relevant behaviour amongst assessors. This understanding can be linked to certain areas related to the performance and organisational requirements of each AC. FOR training should therefore be specifically designed for a certain AC (Lievens, 2002). For example, in an AC where listening skills are observed, the specific listening skills required at the AC will be defined and discussed in detail during the training. A practical example of the listening skills required will then be illustrated or discussed. Certain skills that may be seen as listening skills but are not necessarily required by the particular AC will also be discussed. The aim of this process is to equip the assessors with a mental picture of the competency they will observe during the AC and to eliminate the possibility of assessors using their own mental pictures of how a certain competency manifests. However, Lievens *et al.* (2009) also indicate that research on comprehensive training approaches such as FOR training is lacking.

Schleicher *et al.* (2002) believe that implementing FOR training for assessors can be viewed as an intervention that will have a significant influence on both the construct validity and the criterion validity of ACs. FOR training is recognised as a well-known term in the field of performance appraisal, mostly because of the evidence that FOR has a significant positive effect on assessors' reliability and accuracy (Lievens, 2009; Schleicher *et al.*, 2002). Lievens and Thornton (2005) point out that FOR training not only trains assessors to distinguish between behaviours and dimensions in accordance with a specific framework, but also reduces the cognitive load as a unified scoring framework can be implemented.

Lievens (2002; 2009) and Schleicher *et al.* (2002) claim that if the FOR training approach is followed it should lead to more accurate results by educating assessors to use more effective and appropriate schemas (frames of reference). This argument is supported by the research of various authors, all of whom have reported that FOR training presented higher discriminative validities, criterion validities and rating accuracy (Lievens, 2002; Schleicher *et al.*, 2002). The evidence that FOR trained assessors are better able to use different dimensions accurately (Lievens, 2002), implies that the principles of FOR training should be incorporated into assessor training. The argument of Schleicher *et al.* (2002), that FOR increases overall validity as well as legal defensibility, further emphasises the importance of implementing and following this approach.

After completing an Honours degree in Industrial Psychology a student can register as a psychometrist with the Health Professions Council of South Africa (HPCSA). The HPCSA states that a registered psychometrist should be able to participate in assessment procedures in diverse settings and organisations. Regarding the scope of practice for assessments the HPCSA (Health Professions Council of South Africa, 2010) mentions that during all assessments observers have to declare the limits of their evaluations and that they may not misuse assessment techniques or results. By training graduate psychometrist students in FOR methods, their ability to participate in diverse assessments and settings should be enhanced.

From the discussion above it is clear that by focusing on effective assessor training, more specifically FOR training, construct validity as well as predictive and content validity should increase. It has, however, been speculated that FOR could also influence convergent validity. At the moment, however, there is no conclusive evidence for this speculation. Although international studies exist on assessor training as well as FOR training, no such research exists for training psychometrists in the South African context.

Research objectives

The discussion above leads to the objectives (general and specific) of this research being formulated as follows:

General objective: The general objective of this research was to evaluate a training programme for intern psychometrists being trained as assessors at an assessment centre.

Specific objectives: Specific objectives of this research were:

- To investigate the content and methodology for a frame-of-reference training programme for assessors.
- To evaluate the effects of a frame-of-reference training programme for intern-psychometrists as assessors of an assessment centre.

Method

Research approach

A quantitative design was implemented for this research. The research also fell within the field of experimental research. A classic experimental research design was implemented where two groups were established. According to Salkind (2009) a classic experimental design allows the researcher to extensively explore the effect of the independent variable (FOR training programme) on the dependent variable (participants' knowledge of the subject). Furthermore, a randomised pre-test-post-test control group design was implemented (Salkind, 2009).

Research method

Research participants

The population consisted of postgraduate students at a tertiary institution. Purposive sampling was used to obtain a population of 22 Industrial Psychology students ($N = 22$) who were included in the study. The sample size was governed by data saturation and was determined by the number of participants willing to participate and accessible (Burns & Grove, 1987). The method of purposive sampling is used in incidents where the sampling is not necessarily focused on being random but rather done with a specific outcome in mind and has the goal of providing a sample of information-rich participants (Bryman & Bell, 2011; Maree, 2007; Struwig & Stead, 2007). The participants were predominantly White (91%), Afrikaans speaking (91%) and female (68%). All participants were between the ages of 20 and 25 years.

Measuring instruments

Data was collected by means of ratings of nine competencies of a typical AC simulation. During this process the participants were requested to evaluate independent role-players according to nine competencies whilst viewing a DVD recording of a typical AC simulation. During their evaluation they were asked to award a rating to the role-player on the various competencies. The ratings received from the experimental and the control group were compared and analysed after the pre- and post-test. The effect of the FOR training programme on the participants' practical understanding and their skills in observing behaviour accurately was determined by comparing the results of the pre- and post-tests respectively.

Research procedure

In order to gather data statistically and ethically the research project obtained approval from the university's ethics committee. Once approval was granted, all participants were invited to an information session during which the researches' aims and procedures were explained to them. The participants' consent was obtained, and then they were randomly divided into the control group and the experimental group. This is in accordance with the pre-test-post-test control group design (De Vos, Strydom, Fouché & Delport, 2005). The schedule for the pre-test simulations was then drawn up.

The entire group was subjected to a pre-test assessment. During this pre-test the participants had to evaluate role-players in an AC according to nine predetermined competencies. The experimental group was then subjected to the FOR training programme, whilst the comparison group received no training. The training programme mainly consisted of a series of workshops dedicated to the development of interviewing and assessment skills. The programme was presented by means of two previously recorded ACs during which the participants were taught FOR principles. Once the training programme had been presented, the entire group underwent the post-test. The group again evaluated the DVD recording of the AC that they had seen during the pre-test. The comparison group only underwent the training programme after the post-test had been administered. The ratings of the experimental and control group were compared after the post-test to measure the effect of the FOR training programme. The training period was scheduled for the end of the university's semester, and took place on three consecutive days in order to minimise the carry-over effect.

Statistical analysis

In this study, SPSS (2012) was utilised to determine non-parametric statistics, namely the Mann-Whitney *U*-test and the Wilcoxon Signed Ranks test. The Mann Whitney *U*-test was implemented with the experimental and control groups by comparing the medians, to determine whether the two groups were at the same level prior to the implementation of the FOR training programme. This non-parametric technique is preferred for data measured according to a category or a ranking, as well as for small samples (Pallant, 2010), which was the case in this study. The Wilcoxon Signed Ranks test was then used to determine the difference between the pre- and post-test results in the experimental group. This technique is used with repeated measures, in other words, to measure the participants at two different occasions (Pallant, 2010). Effect sizes were calculated for the results of both the Mann Whitney *U*-test and the Wilcoxon Signed Ranks test. This was done by dividing the *z*-value by the square root of N ($N = 22$). The guidelines set by Cohen (1988) were used to determine the effect size, namely 0.1 = small effect, 0.3 = medium effect and 0.5 = large effect.

Cronbach alpha coefficients were also used to determine the internal consistency and reliability of the ratings received. These statistics were utilised to effectively observe the effect of the training programme on the rating difference and accuracy between the experimental group and control group.

Results

The following section gives an account of the results of the study. Firstly, the content and methodology of the FOR training programme will be reported, and then the Cronbach alpha will be investigated. Finally, the non-parametric statistics will be reported. Table 1 depicts the content and methodology of the FOR training programme.

The training programme was conducted over a three-day period. An existing training programme was adjusted to accommodate frame-of-reference training. From the table it can be seen that the first workshop focused on basic facilitation skills such as listening, objective attending and paraphrasing. The second workshop focused on informing the participants of the principles of an AC and included in-depth discussions of the competencies assessed. During the second day of the training programme, practical exercises were conducted on rating the four candidates (role-players) taking part in two separate one-on-one simulations on the competencies (strategic perspective, interpersonal skills, leadership, conflict management, judgement, self-confidence, assertiveness, persuasive communication and performance under pressure). On the third day of the training programme two workshops were presented. The first one focused on practical exercises for rating the same candidates participating in two different presentation simulations on the discussed dimensions. The second workshop focused on practical exercises to rate two group discussion simulations consisting of four candidates each.

This study consisted of two groups: an experimental group and a control group. In order to answer the second objective of this study, the first step was to determine the internal consistency of the AC (one-on-one, presentation and group discussion simulations) between the experimental group and control group. The results are reported in Table 2.

According to Table 2, the internal consistency for the experimental and control groups for the pre- and post-test is illustrated by reporting the Cronbach alphas. From the table it can be derived that the internal consistency for the experimental group from the pre-test (across all three simulations) ranges between 0.725 and 0.941, and for the post-test between 0.574 and 0.936. Similarly, for the control group the Cronbach alphas for the pre-test (across all three competencies) range from 0.545 to 0.958 and for the post-test between 0.737 and 0.932.

After this the significant differences between the experimental and control groups prior to the FOR training programme were determined in terms of the rating of the nine competencies of the AC. A Mann-Whitney *U*-test revealed no significant differences between the experimental and control groups in their assessment of the one-on-one, presentation and group discussion simulations. For the three simulations utilised in the AC, the Mann-Whitney *U*-test ranged between 31 and 60, the *z*-value ranged between -1.94 and -0.03, the *p*-value ranged between 0.052 and 0.974, and the correlations coefficient ranged between -0.41 and -0.01.

TABLE 1: The content, objectives and methodology of the FOR training programme for assessors at assessment centres.

Workshop	Title	Objective	Method
Day 1			
Session 1	Basic interviewing and facilitation skills	Transferring practical and theoretical knowledge of managing a basic facilitation process	Lecture; Role play
Session 2	Introduction to ACs and competencies	Manifest a comprehension of basic AC principles and practices	Group work; Discussion
		Manifest an understanding of competencies and how to identify them	Lecture
Day 2			
Session 1	Practical work	To observe competencies in role-players' behaviour and evaluate accordingly	Video material; Group discussion
Session 2	Feedback	Provide feedback on evaluations by expert assessors	Video material; Group discussion; Individual coaching session
Day 3			
Session 1	Conclusion	Transferring knowledge	Lecture; Group discussion

TABLE 2: The Cronbach's alphas (α) between the pre- and the post-test for the experimental and control group for the AC.

Cronbach's alpha (α)	Simulation	Candidate	One-on-one	Presentation	Group discussion
Experimental group	Pre-test	1	0.892	0.935	0.941
		2	0.733	0.876	0.908
		3	0.725	0.873	0.876
	Post-test	1	0.912	0.911	0.928
		2	0.813	0.759	0.800
		3	0.936	0.574	0.855
Control group	Pre-test	1	0.883	0.880	0.926
		2	0.958	0.948	0.857
		3	0.871	0.857	0.545
	Post-test	1	0.853	0.810	0.932
		2	0.807	0.927	0.900
		3	0.737	0.808	0.830

The next step was to investigate the difference between the pre- and post-test scores of the experimental group for the nine competencies of the AC (one-on-one, presentation and group discussion simulations). These results are reported in Table 3.

The Wilcoxon Signed Ranks Test indicated a statistically significant reduction in Candidate 3's asssessment of the one-on-one simulation after the FOR training programme, z = -2.81, p = 0.306, with a large effect size (r = -0.60) (see Figure 1). The median score on the aforementioned decreased from the pre-test (Md = 7.56) to the post-test (Md = 6.78). Similarly, a significant reduction was found for Candidate 1 (z = -2.36, p = 0.018, r = -0.50: large effect) and Candidate 3 (z = -2.80, p = .005, r = -0.60: large effect) in the assessment of the group discussion simulation. The median score for Candidate 1 decreased from the pre-test (Md = 6.67) to the post-test (Md = 5.22) and for Candidate 3 from the pre-test (Md = 6.89) to the post-test (Md = 6.22) (See Figure 2).

The differences between the pre- and post-test scores for the control group for the nine competencies of the AC (one-on-one, presentation and group discussion simulations) are reported in Table 4.

Table 4 reveals that, unlike the results for the experimental group, the Wilcoxon Signed Rank Test for the control group shows no statistically significant differences between the pre- and post-test for the AC (one-on-one, presentation, group discussion simulations).

The next step involved the determination of the differences in the rating of the nine competencies of the AC between the experimental and control groups after the FOR training programme had been implemented within the experimental group. A Mann-Whitney U-test revealed a significant difference in the assessment of the one-on-one simulation for Candidate 2 between the experimental group and the control group (U = 25, z = -2.34, p = 0.019, r = -0.50: large effect). Additionally, a significant difference was found in the assessment of the presentation simulation for Candidate 3 between the experimental group and the control group (U = 28, z = -2.14, p = 0.032, r = -0.46: medium effect). The remaining differences between the experimental and control groups were non-significant.

Discussion

This study focused on evaluating a frame-of-reference (FOR) training programme for assessors at an assessment centre. The main aim of the programme was to improve the evaluation and assessment skills of graduate psychometrist students. Generally, the results indicate that the FOR training programme did indeed improve the assessment skills of the experimental group.

TABLE 3: The difference between the pre- and post-test scores for the experimental group for the AC.

Simulation	Pre-test to post-test	z-value	p	R
One-on-one	Candidate 1	-1.02[b]	0.306	-0.22
	Candidate 2	-1.16[c]	0.247	-0.25
	Candidate 3	-2.81[b]	0.005*	-0.60
Presentation	Candidate 1	-1.70[b]	0.090	-0.36
	Candidate 2	-0.87[b]	0.385	-0.19
	Candidate 3	-1.53[b]	0.126	-0.33
Group discussion	Candidate 1	-2.36[b]	0.018*	-0.50
	Candidate 2	-1.65[b]	0.099	-0.35
	Candidate 3	-2.80[b]	0.005*	-0.60

[a], Wilcoxon signed ranks test
[b], Based on positive ranks
[c], Based on negative ranks
*, $p \leq 0.05$
Practically significant correlation: $r \geq 0.10$ (small effect); $r \geq 0.30$ (medium effect); $r \geq 0.50$ (large effect)

TABLE 4: The difference between the pre- and post-test scores for the control group for the AC.

Simulation	Pre-test to post-test	z-value	p	R
One-on-one	Candidate 1	-1.29[b]	0.197	-0.28
	Candidate 2	-0.31[b]	0.755	-0.07
	Candidate 3	-0.98[b]	0.327	-0.21
Presentation	Candidate 1	-1.88[b]	0.060	-0.40
	Candidate 2	-0.15[b]	0.878	-0.03
	Candidate 3	-1.03[b]	0.305	-0.22
Group discussion	Candidate 1	-0.18[b]	0.859	-0.04
	Candidate 2	-0.31[c]	0.759	-0.07
	Candidate 3	-1.65[b]	0.100	-0.38

[a], Wilcoxon Signed Ranks Test
[b], Based on positive ranks
[c], Based on negative ranks
*, $p \leq 0.05$ (there were no values below 0.05)
Practically significant correlation: $r \geq 0.10$ (small effect); $r \geq 0.30$ (medium effect); $r \geq 0.50$ (large effect)

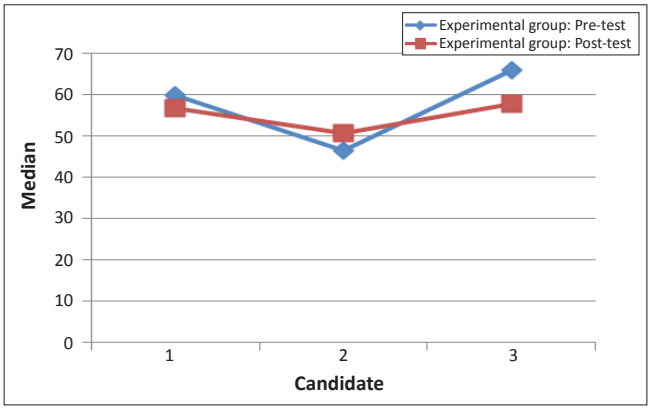

FIGURE 1: The comparison of the pre- and post-test rating for the one-on-one simulation by the experimental group.

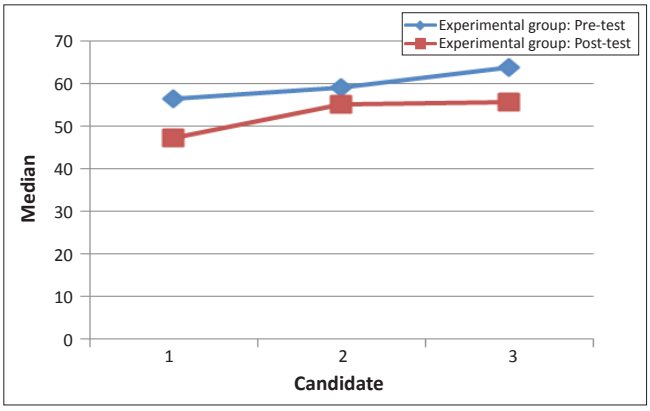

FIGURE 2: The comparison between the pre- and post-test ratings for the group discussion simulation for the experimental group.

According to the Health Professions Council of South Africa (HPCSA), a postgraduate student in the field of Industrial and Organisational Psychology should be able to assist in various assessment procedures in a diversity of settings and organisations. Hence the FOR training programme had the objective of improving the students' basic facilitation skills, their familiarity with the simulations implemented to measure pre-determined competencies as well as their experience and familiarity with the competencies and the assessment process. The results indicated that the FOR training programme increased the participants' familiarity with the one-on-one simulation and partially with the group discussion simulation. However, no significant results could be reported for the participants' ability to assess the presentation simulation.

Literature indicates that an assessment centre is a process that utilises multiple work-like exercises to measure multiple dimensions, pertinent to effective performance in a specific position (Hoffman, Melchers, Blair, Kleinmann & Ladd, 2011; Thornton & Rupp, 2005). Assessors at an assessment centre can be viewed as the individuals who, through multiple work-like exercises, observe, classify and evaluate the multiple dimensions displayed by the candidates taking part in the multiple exercises (Goodstone & Lopez, 2001; Hoffman *et al.*, 2011; International Task Force on Assessment Centre Guidelines, 2010; Schlebusch, 2008). The training of these assessors focuses on developing those abilities that will equip them in the matter of observing, assessing and classifying candidates' behaviour effectively and accurately (Schlebusch, 2008). A technique often used for assessor training is frame-of-reference (FOR) training (Jackson *et al.*, 2009). Frame-of-reference training aims at developing a mutual understanding between the assessors with regard to the competencies required at a specific assessment centre (Lievens & Conway, 2001; Lievens *et al.*, 2009; Schleicher *et al.*, 2002).

The first objective of this study was to investigate the content and methodology of a frame-of-reference training programme for assessors at assessment centres. This resulted in the compilation of a three-day training programme consisting of five separate workshops. An assessor training programme developed by Spangenberg (1997) was adapted for the required context. The first day of the training consisted of two workshops. The first workshop focused on basic facilitation skills such as listening, objective attending and paraphrasing. A helping skills programme providing training in facilitation skills by Du Preez and Jorgensen (in press) was adapted to fit the context and interviewing purposes of the assessor training programme. Research indicates that these are important competencies for assessors at assessment centres (An International Survey of Assessment Centre Practices, 2010).

The second workshop focused on informing the participants of the principles of an AC and provided in-depth discussions of the competencies being assessed. The discussion concerning the principles of an AC focused on the objectives of an AC, the reasons for using an AC, characteristics of an AC, different simulations that can be implemented at an AC and the role and duties of the assessor. The simulations utilised in the training, as well as skills necessary for assessing (observe, record, classify and evaluate or ORCE) were discussed. The competencies utilised in the training were defined and discussed in terms of certain behavioural indicators of these competencies, and the scoring sheet was discussed and explained in detail. The competencies utilised in this study were strategic perspective, interpersonal skills, leadership, conflict management, judgement, self-confidence, assertiveness, persuasive communication, performance under pressure, adaptability, ability to follow instructions, information usage, oral communication and technical and professional knowledge. This corresponds with previous studies on the requirements of a FOR training programme as well as research on the content of typical assessor training programmes (Melchers, Lienhardt, Von Aardburg & Kleinmann, 2011; Bernardin *et al.*, 2000; Schlebusch, 2008; Sulsky & Kline, 2007).

During the second day of the training programme practical exercises were conducted on rating the four candidates (role-players) taking part in two separate one-on-one simulations on the competencies (strategic perspective, interpersonal skills, leadership, conflict management, judgement, self-confidence, assertiveness, persuasive communication and performance under pressure). The contents of this part of the training correspond with studies regarding FOR training (Bernardin *et al.*, 2000; Melchers *et al.*, 2011; Sulsky & Kline, 2007).

On the third day of the training programme two workshops were presented. The first one focused on practical exercises in rating the same candidates taking part in two different presentation simulations on the discussed dimensions. The second workshop focused on practical exercises in rating two group discussion simulations consisting of four candidates each. This part of the training concurs with research on the principles and content of a FOR training programme which point to the importance of providing practical exercises and feedback on these practice ratings (Bernardin *et al.*, 2000; Melchers *et al.*, 2011; Sulsky & Kline, 2007).

Concerning the results of the second objective - to evaluate a frame-of-reference training programme for assessors at assessment centres – the evaluation and assessment skills of the experimental group was found to have improved.

Firstly, the internal consistency for the AC was investigated. The ratings for the competencies of the AC for both the experimental and control group in the pre-test all showed high reliabilities. In the post-test for the experimental group all the competencies also showed high reliabilities. However, the presentation simulation for Candidate 3 had a relatively low reliability for the experimental group. The ratings for the competencies of the AC for the control group in the

pre-test all showed high reliabilities as well. However, the group discussion simulation for Candidate 3 had a relatively low reliability. For the one-on-one simulation, a significant increase in internal consistency was reported for all three candidates. This could indicate that the reliability of the AC had increased with the post-test, which would, in its turn, indicate a positive effect of the FOR training programme. However, no significant increases were reported for the presentation simulation, and a decrease was in fact reported for the presentation simulation during the post-test. The reliabilities for the presentation simulation were still relatively high. This could indicate that although there was a decrease for the post-test, the measurement was still reliable, meaning that out of the three simulations the FOR training programme showed the least improvement for the presentation simulation. The reliabilities for the post-test for this group were, however, all within the acceptable range. This indicates that the experimental group had a good reliability for the pre-test, which made an increase in the post-test even more significant (Lievens, 2002, 2009; Schleicher *et al.*, 2002). This implies that the FOR training led to a more accurate assessment. Various researchers (Jones & Born, 2008; Holmboe, 2004; Lievens, 1998, 2009) have found that assessor expertise is crucial for accurate evaluation. During the FOR training the least attention was given to the presentation simulation. The partially significant results indicated by the group discussion simulation are confirmed in previous research (Melchers, Kleinmann & Prinz, 2010).

The results of the pre-test showed that no significant differences in the rating of the AC existed between the experimental and control groups which could have had an influence on the training programme. This indicates that the two groups were at the same level concerning their knowledge of FOR training prior to the implementation of the programme.

After the experimental group had received the FOR training, the results indicated that the participants improved in their rating of Candidate 2 for the one-on-one simulation. This is an indication that the way in which the participants rated this simulation were similar (the ratings became closer between the participants). The same result was found for the group discussion simulation for two of the three candidates. Furthermore, the results for the control group indicated no statistically significant differences between the pre- and post-test for the AC. Previous studies regarding FOR training support this finding by stating that FOR training promotes a mutual understanding between assessors (Jackson *et al.*, 2005; Lievens, 2001; Lievens *et al.*, 2009; Schleicher *et al.*, 2002; Thornton & Rupp, 2005). It can therefore be concluded that the FOR training programme did in fact create a better mutual understanding and definition of the competencies assessed for each candidate in the experimental group.

A study done by Melchers, Kleinmann and Prinz (2010) found that a group discussion simulation is one of the most difficult simulations to evaluate. Because there are multiple candidates as well as multiple dimensions that have to be evaluated simultaneously, the process may result in cognitive overload. The fact that the experimental group showed the largest improvement in the group discussion simulation could indicate that the FOR training programme had a significant influence on this group.

No constant differences in the ratings of certain competencies were reported during the presentation simulation for the experimental group. One could speculate that a reason could be the fact that during the FOR training, the least amount of time was spent on the rating of the presentation simulation as the participants claimed to feel confident in their rating of its content sooner than with the other simulations. Previous research (Goodstone & Lopez, 2001; Lievens, 2001; Schlebusch, 2008; Hoffman *et al.*, 2011; International Task Force on Assessment Centre Guidelines, 2010) supports the finding that the amount of training for a specific simulation can have an effect on the reliability of the rating for that simulation.

Another observation made during the training on the presentation simulation was that the scenarios used in the training would not necessarily enable an Honours student to truly measure the relevant competencies. Possible explanations for this phenomenon could be that the scenarios for the simulation did not make it possible for the candidates to portray those competencies seen as necessary for an Industrial Psychology Honours student. The scenarios required specific expertise in subjects which are not related to Industrial Psychology. The presentation was also not delivered in front of an audience but only recorded. This made it difficult for candidates to portray the industrial-specific competencies required for the simulation. Therefore there was some difficulty to truly practise the ratings for Honours student competencies.

The results further indicated significant differences for two simulations between the experimental group and the control group from the pre-test to the post-test, namely in the one-on-one simulation for Candidate 2, and of the presentation simulation for Candidate 3. This implies that the two groups measured differently on only two simulations (ideally they should measure differently on more simulations; the experimental group therefore showing a significant improvement compared to the control group). It is possible that these results can be explained as a stochastic error. Kahane (2008, p. 218) describes a stochastic error as 'variables or processes that are inherently random (i.e. not deterministic or exact)'. A similar finding was reported by Melchers *et al.* (2011).

Overall, there is an indication that there was a better mutual understanding of the competencies in the experimental group after the FOR training. This indicates that there was indeed improvement in the assessment skills of the graduate Industrial Psychology students at assessment centres. In conclusion it can be stated that the FOR training programme

had an effect on the reliability of the ratings awarded by the assessors. As stated in previous studies, accurate rating by assessors could have a significant effect on the construct validity of the AC, and therefore this FOR training could have positive and practical implications for various AC processes.

Limitations

With regard to the limitations of the present study, the following can be said: Firstly, the length of the training should be extended so that participants can gain more practical experience, enabling them to rate efficiently and accurately on all nine competencies. Although the current programme consisted of three days of intensive training, the cognitive load was immense. Secondly, the participants were postgraduate students in Industrial Psychology and, although this kind of training did not form part of their course work, they had previous, albeit limited, knowledge of assessment centres. This could have had an effect on the experiment in that the pre-test results were predominantly positive and it was therefore difficult to prove significant differences in the post-test results. Lastly, the sample size (N = 22) could be seen as a possible limitation of this study, as in such a small sample size, a single irregular rating can influence the interpretation of the results. However, the entire population available was utilised and a larger sample size was not possible.

Conclusion

Despite these limitations, the research findings have important implications for future research. One suggestion for future research on a FOR training programme would be to use a population group without any prior behavioural science training. The International Survey of Assessment Centre Practices (2010) shows that although HR members of staff are mostly used as assessors, line managers comprise 53% and members of staff with expertise 27% of assessors used.

Another recommendation for future research is to consider the design of the AC simulations being used. A better design together with an effective training programme for assessors could improve the construct validity of an AC significantly.

Acknowledgements

The authors wish to acknowledge funding support from the National Research Foundation (NRF). Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and therefore the NRF does not accept any liability in regard thereto.

Competing interests

The authors declare that they have no financial or personal relationships which may have inappropriately influenced them in writing this article.

Authors' contributions

G.M. (North-West University) conducted the data collection, assisted with the data analysis, and wrote the manuscript. L.I.J. (North-West University) supervised the study and assisted with the data collection and the development of the manuscript. J.A.N. (North-West University) assisted with the data analysis and prepared manuscript for review and subsequent publication. D.M. (University of Pretoria) co-supervised the study and commented on the final draft of the manuscript.

References

An International Survey of Assessment Centre Practices (2010). *The global research questionnaire.* Surrey, UK: Assessment & Development Consultants Ltd.

Arthur, W., Day, E., McNelly, T.L., & Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154. http://dx.doi.org/10.1111/j.1744-6570.2003.tb00146.x

Bernardin, H.J., Buckley, M.R., Tyler, C.L., & Wiese, D.S. (2000). A reconsideration of strategies for rater training. *Research in Personnel and Human Resources Management, 18*, 221–274.

Bryman, A., & Bell, E. (2011). *Business research methods.* Oxford, UK: Oxford University Press.

Burns, N., & Grove, S.K. (1987). *The practice of nursing research.* Philadelphia: WB Saunders.

Cohen J. (1988). *Statistical power analysis* (2nd edn.). Hillsdale: Erlbaum.

Collins, J.M., Schmidt, F.L., Sanchez-Ku, M., Thomas, L., McDaniel, M.A., & Le, H. (2003). Can basic individual differences shed light on the construct of assessment centre evaluations? *International Journal of Selection and Assessment, 11*(1), 17–29. http://dx.doi.org/10.1111/1468-2389.00223

De Vos, A.S., Strydom, H., Fouche, C.B., & Delport, C.S.L. (2005). *Research at grass roots for the social sciences and human services professions* (3rd edn.). Pretoria, South Africa: Van Schaik Publishers.

Dilchert, S., & Ones, D.S. (2009). Assessment centre dimensions: Individual differences correlates and meta-analytical incremental validity. *International Journal of Selection and Assessment, 17*(3), 254–270. http://dx.doi.org/10.1111/j.1468-2389.2009.00468.x

Du Preez, J. & Jorgensen, L.I. (in press). The evaluation of a helping skills training programme for intern-psychometrists. *Journal of Psychology in Africa*.

Eurich, T.L., Krause, D.E., Cigularov, K., & Thornton, G.C. III. (2009). Assessment centres: Current practice in the United States. *Journal of Business & Psychology, 24*, 387–407. http://dx.doi.org/10.1007/s10869-009-9123-3

Goodstone, M.S., & Lopez, F.E. (2001). The frame of reference approach as a solution to an assessment centre dilemma. *Consulting Psychology Journal: Practice and Research, 53*(2), 96–107. http://dx.doi.org/10.1037/1061-4087.53.2.96

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah: Lawrence Erlbaum Associates.

Health Professions Council of South Africa (2010). Health Professions Act 1974 (Act 56 of 1974). Retrieved June 08, 2011, from http://www.hpcsa.co.za

Hoffman, B.J., Melchers, K.G., Blair, C.A., Kleinmann, M., & Ladd, R.T. (2011). Exercises and dimensions are the currency of assessment centres. *Personnel Psychology, 64,* 351–395. http://dx.doi.org/10.1111/j.1744-6570.2011.01213.x

Holmboe, E. (2004). Faculty and the observation of trainees' clinical skills: Problems and opportunities. *Academic Medicine, 79*(1), 16–22. http://dx.doi.org/10.1097/00001888-200401000-00006, PMid:14690992

International Task Force on Assessment Centre Guidelines (2010). *Guidelines and ethical considerations for assessment centre operations.* San Francisco: Development Dimensions International.

Jackson, D.J.R., Atkins, S.G., Fletcher, R.B., & Stillman, J.A. (2005). Frame of reference training for assessment centres: Effects on inter-rater reliability when rating behaviours and ability traits. *Public Personnel Management, 34*(1), 17–30.

Joiner, D.A. (2004, June). *Assessment centre trends: Assessment centre issues and resulting trends.* Paper presented at the 28th annual meeting of IPMAAC, Seattle, Washington.

Jones, R.G., & Born, M.P. (2008). Assessor constructs in use as the missing component in validation of assessment centre dimensions: A critique and directions for research. *International Journal of Selection and Assessment, 16*(3), 229–238. http://dx.doi.org/10.1111/j.1468-2389.2008.00429.x

Kahane, L.H. (2008). *Regression basics* (2nd edn.). Thousand Oaks: Sage.

Krause, D.E., & Gebert, D. (2003). A comparison of assessment centre practices in organisations in German speaking regions and the United States. *International Journal of Selection and Assessment, 11,* 297–312. http://dx.doi.org/10.1111/j.0965-075X.2003.00253.x

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152. http://dx.doi.org/10.1111/1468-2389.00085

Lievens, F., & Conway, J.M. (2001). Dimension and exercise variance in assessment centre scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*(6), 1202–1222. http://dx.doi.org/10.1037/0021-9010.86.6.1202, PMid:11768062

Lievens, F. (2002). An examination of the accuracy of slogans related to assessment centres. *Personnel Review, 31,* 86–102. http://dx.doi.org/10.1108/00483480210412436

Lievens, F., & Thornton, G.C. III. (2005). Assessment centres: Recent developments in practice and research. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell Handbook of Personnel Selection* (pp. 243–264). Malden: Blackwell.

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises and dancing bears. *European Journal of Work and Organisational Psychology, 18*(1), 102–121. http://dx.doi.org/10.1080/13594320802058997

Lievens, F., Tett, R.P., & Schleicher, D.J. (2009). Assessment centres at the crossroads: Toward a reconceptualization of assessment centre exercises. *Research in Personnel and Human Resources Management, 28,* 99–152. http://dx.doi.org/10.1108/S0742-7301(2009)0000028006

Maree, K. (2007). *First steps in research.* Pretoria, South Africa: Van Schaik Publishers.

Meiring, D. (2008). Assessment centres in South Africa. In S. Schlebusch, & G. Roodt (Eds.), *Assessment Centres: Unlocking potential for growth* (pp. 21–31). Randburg, South Africa: Knowres Publishing.

Melchers, K.G., Kleinmann, M., & Prinz, M.A. (2010). Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment centre candidates on rating quality. *International Journal of Selection and Assessment, 18*(3), 329–341. http://dx.doi.org/10.1111/j.1468-2389.2010.00516.x

Melchers, K.G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64*, 53–87. http://dx.doi.org/10.1111/j.1744-6570.2010.01202.x

Pallant, J. (2010). *SPSS Survival Manual. A step by step guide to data analysis using SPSS* (4th edn). Berskhire, England: McGraw-Hill

Pell, G., Homer, M.S., & Roberts, T.E. (2008). Assessor training: Its effect on criteria-based assessment in a medical context. *International Journal of Research & Method in Education, 31*(2), 143–154. http://dx.doi.org/10.1080/17437270802124525

Salkind, N.J. (2009). *Exploring research* (pp. 243–251). Saddle River: Pearson Education.

Schlebusch, S. (2008). Before the centre. In S. Schlebusch & G. Roodt (Eds.), *Assessment Centres: Unlocking potential for growth* (pp. 176–196). Randburg, South Africa: Knowres Publishing.

Schleicher, D.J., Day, D.V., Mayes, B.T., & Riggio, R.E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centres. *Journal of Applied Psychology, 87*(4), 735–746. http://dx.doi.org/10.1037/0021-9010.87.4.735, PMid:12184577

Spangenberg, H.H., (1997). *Middle management development centre. Training Manual.* Pretoria, South Africa: Author.

SPSS Inc. (2012). *IBM SPSS 20.0 for Windows.* Chicago: SPSS Incorporated.

Struwig, F.W., & Stead, G.B. (2001). *Planning, designing and reporting research.* Cape Town, South Africa: Pearson Education South Africa.

Sulsky, L.M., & Kline, T.J.B. (2007). Understanding frame-of-reference training success: A social learning theory perspective. *International Journal of Training and Development, 11*(2), 121–131. http://dx.doi.org/10.1111/j.1468-2419.2007.00273.x

Thornton, G.C., III., Murphy, K.R., Everest, T.M., & Hoffman, C.C. (2000). Higher cost, lower validity and higher utility: Comparing the utilities of two tests that differ in validity, costs, and selectivity. *International Journal of Selection and Assessment, 8,* 61–75. http://dx.doi.org/10.1111/1468-2389.00134

Thornton, G.C., III., & Mueller-Hanson, R. (2004). *Developing organisational simulations: A guide for practitioners and students.* Mahwah: Lawrence Erlbaum.

Thornton, G.C., & Rupp, D.E. (2006). *Assessment centres in human resource management: Strategies for prediction, diagnosis and development.* Mahwah: Lawrence Erlbaum.