Buckett, A. et al. (2017). General performance factors and group differences in assessment center ratings.
Journal of Managerial Psychology, 32(4): 298-313.
https://doi.org/10.1108/JMP-08-2016-0264

General performance factors and group differences in assessment center ratings

Anne Buckett, Jürgen Reiner Becker and Gert Roodt

Abstract

Purpose – The purpose of this paper is to establish the extent of general performance factors (GPF) in assessment center (AC) exercises and dimensions. The study further aims to determine if larger GPF contributes to larger ethnic group differences across exercises and dimensions that are more cognitively loaded in an emerging market context.
Design/methodology/approach – The authors analyzed data across three independent AC samples (Sample 1: N = 172; Sample 2: N = 281; Sample 3: N = 428). The Schmid-Leiman solution was used to determine the extent of GPF in AC exercises and dimensions. An independent samples t-test and Cohen's d was used to determine the size of ethnic group differences across exercises and dimensions.
Findings – The results indicate that GPF is consistently large for the in-basket exercise. Furthermore, dimensions that are more cognitively loaded, such as problem solving, strategic thinking, and business acumen, seem to produce the largest ethnic group differences. Overall, the research indicates that larger GPF is associated with larger ethnic group differences in relation to specific AC dimensions and exercises.
Originality/value – The authors add to the literature by investigating the prevalence of a GPF in AC ratings across AC exercises and dimensions. A novel contribution of the research attempts to link the prevalence of a GPF in AC ratings to group membership in South Africa. The study offers an alternative statistical analysis procedure to examine GPF in AC ratings.

Introduction

Organizational psychology researchers have, for decades, attempted to confirm if dimensions or exercises are the main components being measured in the assessment center (AC) (Lance, 2008). AC scholars have subsequently arrived at a blended solution to address this issue, with research demonstrating that dimensions and exercises are two central components of the AC method (Hoffman et al., 2011; Melchers et al., 2012). Furthermore, contemporary research has identified a third component measured in the AC, namely, a general performance factor (GPF) (Lance et al., 2004).

Research suggests that ACs tend to produce smaller ethnic group differences than other recognized predictors (Melchers and Annen, 2010). Although the presence of group differences between Blacks and Whites in psychometric assessments is well established

(e.g. Rushton et al., 2003), there has been less focus on group membership in relation to ACs. Internationally, efforts to reduce the emphasis on group differences inherent in psychometric assessments have resulted in a number of alternative procedures and predictors. These include within-group norming, separate cut-off scores for different groups, and using alternative measures, such as ACs (Ployhart and Holtz, 2008).

However, two notable large-scale reviews suggest that the perception that ACs do not indicate group differences may be misplaced and that these differences in AC findings are, in fact, larger than previously reported in the literature (Bobko and Roth, 2013; Dean et al., 2008). One possible explanation for these differences could be due to a GPF. Although GPF is well documented in the job performance literature (Viswesvaran et al., 2005), it is relatively new in AC research. Viswesvaran et al. found that a GPF in job performance ratings could be partially explained by contextual performance and certain individual differences, such as cognitive ability and personality traits. Studies of a similar nature in AC research have shown that a GPF correlates with cognitive ability, a factor contributing to group differences in psychometric assessments (Hoffman, 2012; Hoffman et al., 2011).

As an emerging market economy, South African organizations are faced with slow labor market reform, diverse demographics and disparate opportunities for economically active populations (Euromonitor Research, 2011; Faulk and Salem, 2014). Using ACs as an alternative to overcome some of these challenges seems to be a practical solution. ACs are regularly used for managerial selection and development in South Africa (Krause et al., 2011). However, there is limited local published research regarding the impact of a GPF in AC ratings and on differences in candidate performance related to group membership. Our research thus attempted to establish the existence of a GPF in AC ratings. The research was based on the premise that the presence of a GPF may lead to significant group differences similar to the differences found in standardized psychometric tests.

To further prove this premise, we wanted to establish whether group differences in ACs were more pronounced on cognitively demanding exercises and dimensions. In particular, the research aimed to establish whether specific AC exercises and/or specific dimensions showed more variation between groups due to the presence of a GPF that could be attributed to cognitive ability (Dean et al., 2008; Goldstein et al., 1998, 2001).

The study thus makes a contribution to the literature in five ways. First, limited research has attempted to link the prevalence of a GPF in AC ratings to group membership. Second, the research on ethnic group differences in AC dimensions is limited in the literature. Third, there is little published research on the prevalence of a GPF across exercise types. Fourth, the context of the study from an emerging market economy makes a novel contribution to the extensive literature on ACs predominantly from North America and Western Europe. Finally, a unique factor analytic technique was used, namely, the Schmid-Leiman solution (SLS), to examine the relative size of a GPF across exercises.

Psychometric testing and group differences

The use of psychometric assessments for the purpose of selection and development has a rich history in South Africa, yet continues to receive mixed and even negative reactions from test users (Theron, 2007). Prior to 1994 psychometric assessment research and application predominantly focused on the White population, leading to a perception that psychometric assessments are reflective of White culture and values, and biased toward other groups (Foxcroft and Roodt, 2005; Meiring, 2007).

The unbalanced application of psychometric assessments preceding 1994 and the disproportionate impact on access to job opportunities resulted in the promulgation of legislation to mitigate the effects of past inequitable assessment practices. Specifically, the Republic of South Africa (1998) regulates the use of psychological assessments and proposes three criteria to govern their use. First, assessments must be scientifically valid and reliable. Second, assessments must be applied in a fair manner to employees. Third, assessments must not be biased against any specific group.

Despite the presence of legislation governing assessment practises, the finding of ethnic group differences on psychometric assessments remains an inescapable reality for decision makers in South African organizations (Meiring, 2007). This results in apprehension regarding the use of psychometric assessments as organizations need to balance competing goals in the exchange between efficiency (i.e. using the most appropriate test) and equality (i.e. increasing workplace diversity while reducing bias and adverse impact in predictor measures) (Theron, 2007).

Against this backdrop, a plethora of local research on psychometric assessments has ensued (Meiring et al., 2005). Findings from these studies are generally mixed; resulting in wide-ranging recommendations such as eliminating testing (Nell, 2000), using locally designed tests (Meiring et al., 2007), and replacing traditional tests with ACs (Charoux, 1997).

Research on ACs in South Africa in the early 1990s focused extensively on the fairness of the method (Kriek et al., 1994) and the ability to predict Black leadership potential (Charoux and Hurst, 1992). However, none of the referenced studies investigated group differences across AC exercises and dimensions and in relation to a GPF. The sensitivity to group differences may therefore be more pronounced in South Africa, and possibly similar emerging markets, due to the unique socio-political history of the country and the legislative framework. This study therefore addresses an important issue in South Africa given the limited published research into group differences in AC ratings in this context.

ACs and group differences

The AC is a method that uses behavioral simulation exercises, such as role play (RP), group, and in-basket (IB) exercises, to replicate on-the-job performance of participants for selection or development purposes (Thornton et al., 2015). The method is multi-layered and consists of at least two simulation exercises that are mapped to job-relevant dimensions (Meiring and Buckett, 2016).

ACs have historically been regarded as useful predictors of job performance that tend to produce smaller group differences than other recognized predictors (Melchers and Annen, 2010). Although topical AC research has highlighted group differences being as high as ds = 0.56 (Bobko and Roth, 2013) ACs continue to retain a positive reputation as a more predictive method due to the close alignment of AC exercises to the job and confirmed evidence of criterion-related validity (Krause et al., 2006).

However, closer inspection of AC exercises and dimensions suggests that group differences might be more pronounced when cognitive ability is involved (Goldstein et al., 1998; Goldstein et al., 2001). There appears to be a clear distinction between different types of AC exercises and their level of cognitive complexity (Goldstein et al., 1998). AC exercises containing more written content and language comprehension produce the largest group differences (e.g. IB exercise: ds = 0.35). Conversely, AC exercises requiring less information processing skills and more interaction with people result in lower group differences (e.g. RP exercise: ds = 0.03). Group differences are also present in relation to the cognitive load of dimensions (Goldstein et al., 2001). Differences appear to be larger on dimensions that are cognitively loaded, such as problem solving, strategic thinking, and planning and organizing (PO). Group differences are reported to be lower on interpersonal dimensions such as interpersonal skills and empathy. The finding of group differences in relation to the cognitive complexity of the AC exercise and dimension is analogous to the results reported in cognitive ability literature (Schmidt and Hunter, 1998). We therefore expect to see larger group differences in relation to specific AC exercises and dimensions:

H1a. The IB exercise will demonstrate larger group differences than the RP exercise and group exercise.
H1b. Cognitively loaded dimensions will demonstrate larger group differences than interpersonal dimensions.

GPF in ACs

AC researchers have dedicated decades to examining the internal structure of ACs. Specifically, the literature has attempted to find support for either dimensions (Bowler and Woehr, 2006) or exercises (Jackson et al., 2010) as the focal constructs measuring performance in ACs. The culmination of this research focus has resulted in the view that both dimensions and exercises contribute toward the underlying structure of ACs (Hoffman et al., 2011; Putka and Hoffman, 2013). Therefore, AC performance is best interpreted as an interaction between candidates, dimensions, exercises, and the situation being assessed (Hoffman, 2012). This interpretation makes sense because ACs are designed to sample elements of the job and job performance is likely to vary depending on the person, the situation and moderating contextual variables, such as knowledge, skills, and cognitive ability.

An outcome of this research has identified that a GPF forms part of the internal structure of AC ratings (Hoffman et al., 2011). In AC research, the presence of a GPF could indicate the general effectiveness of a participant across dimensions and exercises (Thornton et al., 2015). Kuncel and Sackett (2014) provide three additional explanations for a GPF in AC ratings.

First, GPF could indicate a set of effective behaviors that are relevant across most situations. Second, participants that can identify the criteria being measured generally perform better in the AC. Third, a degree of test familiarity with simulations could contribute to a GPF. GPF therefore provides a situational explanation for additional variance across AC ratings that is not accounted for purely by dimensions and exercises (Thornton et al., 2015).

The actual components of the GPF in AC ratings are, as yet, not decided. In addition to accounting for the general effectiveness of participants, research on the GPF in AC ratings indicates, for example, a relationship to cognitive ability (Hoffman et al., 2011) and conscientiousness (Lance et al., 2007). These findings resonate with research demonstrating that cognitive ability predicts candidate performance in AC exercises (Hoffman et al., 2015) and AC dimensions (Meriac et al., 2014). We can therefore assert that the presence of a GPF in AC ratings might reflect cognitive ability to some extent:

H2a. Variance explained by the GPF will be proportionately larger in the IB exercise than the RP exercise and group exercise.

Method

Three independent samples were used in the current study. One sample was from a private organization and two samples were from a government organization. Participants were job incumbents who were assessed for development purposes. Simulation exercises were designed for the two organizations using dimensions specified by the organization with particular development objectives in mind. There were five common features across the ACs. First, the exercises were pretested prior to implementation. Second, only expert assessors were used. Assessors were diverse in terms of ethnicity. Third, each participant was rated by a single assessor. It was therefore not possible to measure interrater reliability as only one assessor rated a participant at a given time. Scoring was done using the within-exercise approach whereby an assessor rated all the dimensions of a specific exercise for one participant at the end of the exercise. Fourth, frame-of-reference training was provided to the assessors prior to the roll-out of the first AC and consisted of a practical component and background on the organization. Finally, the rating scale was a four-point scale where 1: major development, 2: minor development, 3: competence, and 4: a strength.

Sample 1 participants

Participants were 172 team leaders working in a private sector organization based in the renewable energy industry in South Africa[1]. The sample consisted of 44 percent Blacks and 37 percent Whites. Two-thirds of the group was male (66 percent male, 34 percent female) with a mean age of 39 years. Participants were assessed for development purposes and completed a one-day AC.

Sample 1 AC procedure

The AC consisted of three simulation exercises designed to measure five dimensions. The three exercises included: a group exercise: four to six participants working together to address five work-related management problems; a RP exercise: a counseling discussion

with a non-performing subordinate; and an IB exercise: dealing with a range of issues in writing. The dimensions were business acumen (BA), communication (C), fostering relationships (FR), leadership (L), and results driven (RD). BA, L, and RD were measured in all the exercises. C and FR were measured in the group exercise and the RP exercise but not in the IB exercise.

Participants were assessed in groups of 4-12. The rating forms were designed to measure three to five dimensions per exercise. In total, 7-16 behavioral indicators per dimension were provided in the rating forms. Participants received individual feedback on their AC performance at the end of the day. Data were collected over the course of three years from 2012 to 2014[2]. The same group of assessors was used for the duration of the project.

Sample 2 participants

Participants were 281 team leaders working for a government department in South Africa. The sample consisted of 56.6 percent Blacks and 31.3 percent Whites. The gender composition for this group was predominantly female (58.8 percent female, 41.3 percent male). The age of participants was not reported. Participants were assessed for development purposes and completed a half-day AC.

Sample 2 AC procedure

The AC consisted of two simulation exercises and a competency-based interview, designed to measure eight dimensions. The exercises included: an IB exercise: participants had to address a range of supervisory issues in writing; and a RP exercise: participants had to deal with a non-performing subordinate. The dimensions measured across the exercises included team management (TM), customer service orientation, communication (C), interpersonal interaction (INT), change orientation, self-management, PO, and problem analysis and decision making (PS). Three dimensions (i.e. TM, INT, and PS) were measured multiple times in the AC.

Participants were assessed in groups of 8-16. An average of five behavioral indicators was listed under each dimension. Reports were generated using an automated template and feedback took place toward the end of the project. Data were collected over the course of six months.

Sample 3 participants

Participants were 428 middle managers working for a government department. The sample consisted of 56.3 percent Blacks and 32.3 percent Whites. The gender composition for this group was predominantly male (58.4 percent male, 41.6 percent female). The age of participants was not reported. Participants were assessed as part of the same development project described for Sample 2 and completed a half-day AC.

Sample 3 AC procedure

The AC consisted of two simulation exercises and a competency-based interview, designed to measure eight dimensions. The exercises included: an IB exercise: participants had to address a range of middle management issues in writing; and a RP exercise: the participant

had to work with a colleague to narrow down a selection of projects for the organization. The dimensions measured across the exercises included applied strategic thinking (AST), PS, developing others, impact and influence, customer focus and responsiveness, managing interpersonal conflict and resolving problems (MIC), networking and building bonds, PO, and team leadership. The same procedure described in Sample 2 was followed for feedback and reporting. Four dimensions (i.e. AST, PS, MIC, and PO) were measured multiple times in the AC.

Table I provides the dimension by exercise matrix across the three samples. Each dimension has been linked to Arthur et al. (2003) taxonomy.

Common Dimension Taxonomy (Arthur *et al.*, 2003)	Associated Sample Dimensions	Sample 1 GE	Sample 1 RP	Sample 1 IB	Sample 2 RP	Sample 2 IB	Sample 3 RP	Sample 3 IB	Generic Definitions as defined by Arthur *et al.* (2003, pp. 134-136)
Problem solving	BA; PS; AST	✓	✓	✓	✓	✓	✓	✓	The extent to which an individual systematically gathers information; understands relevant technical and professional information; effectively analyzes data and information; generates viable options, ideas, and solutions; selects appropriate courses of action for problems and situations; uses available resources in new ways; and generates and recognizes imaginative solutions
Communication	C	✓	✓						The extent to which an individual conveys oral and written information and responds to questions and challenges
Consideration/ Awareness of others	FR; INT; MIC	✓	✓		✓	✓	✓	✓	The extent to which an individual's actions reflect a consideration for the feelings and needs of others as well as an awareness of the impact and implications of decisions relevant to other components both inside and outside the organization
Influencing others	L; TM	✓	✓	✓	✓	✓			The extent to which an individual persuades others to do something or adopt a point of view in order to produce desired results and takes action in which the dominant influence is one's own convictions rather than the influence of others' opinions
Drive	RD	✓	✓	✓					The extent to which an individual originates and maintains a high activity level, sets high performance standards and persists in their achievement, and expresses the desire to advance to higher job levels
Organizing and planning	PO						✓	✓	The extent to which an individual systematically arranges his/her own work and resources as well as that of others for efficient ask accomplishment; and the extent to which an individual anticipates and prepares for the future

Notes: Sample 1: BA, business acumen; C, communication; FR, fostering relationships; L, leadership; RD, results driven. Sample 2: TM, team management; INT, interpersonal interaction; PS, problem analysis and decision making. Sample 3: AST, applied strategic thinking; PS, problem analysis and decision making; MIC, managing interpersonal conflict and resolving problems; PO, planning and organizing; GE, group exercise; RP, role play exercise; IB, in-basket exercise

Table I.
Dimension *x* exercise matrix across three samples and correspondence with common dimension taxonomy

Statistical analysis

Two statistical procedures were applied to each data set. First, to determine the extent of GPF in AC exercises and dimensions the SLS syntax was applied to the data (Wolff and Preising, 2005). The purpose of this analysis was to obtain a better understanding of the factor structure of AC ratings by identifying the relationship between items and higher-order factors. With this procedure, first-order factor analysis (FA) was conducted after which the SLS syntax is applied to determine a higher-order factor structure. First-order and second-order loadings are orthogonolized in order to see the pattern of salient factor

loadings more clearly between manifest items and the hierarchical latent factors. There are two primary advantages of the SLS over the confirmatory factor analysis (CFA) approach, which is typically applied to AC ratings to determine factorial relationships. First, the analysis shows the relative contribution of each behavioral indicator in relation to the first- and second-order factors. Second, the analysis is not plagued by issues of non-convergence endemic to hierarchical CFA analyses (Wolff and Preising, 2005).

Second, to examine group differences in AC ratings, the Black-White difference was calculated and an independent samples t-test was conducted whereby the dependent variable was overall exercise score and a dichotomous independent variable was used to represented ethnicity (Pallant, 2013). The results from this analysis were then used to determine the effect sizes with Cohen's d (Lakens, 2013). This procedure was repeated per exercise. The same approach was followed to examine group differences on the dimension scores. In this stage of analysis, the dependent variables were the final dimension ratings (FDRs). Thus, in the first round of analyses, the goal was to explain the group differences by investigating the prevalence and magnitude of the GPF in exercise and dimension ratings. In the second round of analyses, the goal was to investigate the size of group differences across exercise and dimension ratings.

Results

The results will be reported on two levels. First, the presence of a GPF in relation to the general factors that can be extracted by the SLS across exercises and dimensions for the three samples and second, results pertaining to the presence and size of group differences across exercises and dimensions will be presented. Although past research has modeled general performance as first-order factors in CFA matrices (e.g. Hoffman et al., 2011), we chose to specify general performance as a second-order factor composed of first-order latent dimension variables. This approach is desirable in the present context for three reasons. First, AC structure by design is hierarchical and the SLS can identify the structural relations between different factors. Second, the SLS allows for direct comparison between second-order factors and first-order factors. This enables richer factor interpretation. Third, the SLS can be used to identify the independent total impact of factors, in other words, the proportion of variance attributed to each factor (Wolff and Preising, 2005).

SLS

The goal of the SLS was to estimate the magnitude of GPF in exercise and dimension ratings. The SLS analysis was completed in three phases. In Phase 1, first-order FA was carried out by means of principal axis FA for each AC exercise across the three samples. FA was constrained to three factors for each exercise across the three samples, with the exception of the IB in Sample 2 which was constrained to two factors, since extra factors did not contribute significant amounts of additional variance (Hayton et al., 2004). After the first iteration, items with factor loadings o 0.3 and items that were cross-loaded were removed and FA was conducted again. This procedure was followed until item issues were resolved and a solid factor solution was returned. A second-order factor was extracted for each exercise across the three samples in Phase 2. Second-order principal FA indicates what general factors are primarily responsible for the

correlations between the first-order factors. The second-order principal axis FA confirms the relationship between the factors in each exercise across the three samples.

In Phase 3, the second-order factors are transformed by the SLS syntax to directly determine the influence of general factors on the underlying factor structure across exercises and dimensions. Table II reports the percentage of extracted variance that can be explained by GPF and first-order factors across the three samples (H2a).

H2a was supported as the GPF accounts for 66.5-78.2 percent of the variance explained in the IB across the three samples. This suggests that a large proportion of the variance in these exercises is attributable to general factors rather than first-order factors. A similar pattern is observed for GPF in the RP in Sample 2 (73.9 percent) and Sample 3 (69.8 percent). However, the pattern for the RP in Sample 3 is reversed and 68 percent of the extracted variance can be explained by the three first-order factors (i.e. dimensions) while GPF accounts for 32 percent of the variance (i.e. general factors). With the exception of the RP in Sample 1, GPF accounts for 58.2-78.2 percent of the variance explained in the exercises. Samples 2 and 3 show higher percentages of variance that can be explained by general factors across all exercises while Sample 1 shows more variation in results across the exercises.

We attempted to apply the SLS to the dimensions. In order to run the SLS two or more factors need to be extracted in Phase 1. Since the FA suggested that the structure of the dimensions was unidimensional it was not possible to extract an additional performance factor. We found that, in relation to the dimensions as the primary unit of analysis in the SLS, only one factor was extracted in Phase 1 so the analysis could not be completed.

Independent samples t-test and Cohen's d

An independent samples t-test was conducted to explore the impact of ethnicity on overall exercise scores and dimensions in the AC (H1a-H1b). To interpret the strength of the different effect sizes when comparing Blacks and Whites Cohen's d were calculated. The guidelines for interpreting this value are 0.2 (small), 0.5 (medium), and 0.8 (large) (Cohen, 1988). At this point, only dimensions that were fully crossed with exercises were included to report intercorrelations.

The results of the independent samples t-test exploring the impact of ethnicity on overall exercise scores are reported in Table III across the three samples (H1a).

H1a was supported as there was a statistically significant difference between Blacks and Whites at the 0.01 significance level for all the AC exercises across the three samples. The magnitude of the differences was moderate for GE (Sample 1: d = 0.38), RP (Sample 1: d = 0.53; Sample 2: d = 0.66; Sample 3: d = 0.34), and IB (Sample 2: d = 0.73; Sample 3: d = 0.56). The effect size between groups for IB (d = 0.95) in Sample 1 was large. The intercorrelations between the AC exercises across the three samples were small to moderate.

An independent samples t-test was conducted to explore the impact of ethnicity on overall dimension scores in the AC (H1b). Tables IV-VI present the descriptive statistics for

the dimensions across Samples 1-3, respectively. The tables contain means, standard deviations, Cohen's d, and intercorrelations.

H1b was partially supported. There was a statistically significant difference between Blacks and Whites at the 0.01 level for all of the dimensions across the three samples, with the exception of Sample 1. Dimension C (Sample 1) was the only dimension that did not yield a statistically significant difference. In all the independent samples t-tests, the mean dimension ratings were higher for Whites than for Blacks. This result is further confirmed by the large effect sizes in Sample 1 for BA (d = 1.00), L (d = 0.94), and RD (d = 0.94); and Sample 2 for PS (d = 0.8). In contrast to the findings in the other two samples, the magnitude of the differences in Sample 3 for similar dimensions was smaller than expected. The effect size for PO (d = 0.52) was moderate and small for MIC (d = 0.30), PS (d = 0.47), and AST (d = 0.49). The intercorrelations between the dimensions ranged from moderate to large.

	Sample 1						Sample 2						Sample 3					
	IB		RP		GE		IB		RP				IB		RP			
	H2	%	H2	%	H2	%	H2	%	H2	%			H2	%	H2	%		
GPF	9.941	74.50	3.419	32.00	6.830	58.20	7.594	78.20	9.532	73.90			6.564	66.50	7.254	69.80		
F1	1.747	13.10	3.555	33.20	1.800	15.30	1.095	11.30	1.433	11.10			1.283	13.00	1.324	12.70		
F2	0.594	4.50	2.011	18.80	1.453	12.40	1.018	10.50	0.646	5.00			1.046	10.60	0.841	8.10		
F3	1.055	7.90	1.716	16.00	1.650	14.10			1.280	9.90			0.977	9.90	0.979	9.40		
Total	13.337	100	10.700	100	11.733	100	9.706	100	12.891	100			9.870	100	10.398	100		
% of extracted variance explained by general factors		74.5		32		58.2		78.20		73.90				66.50		69.80		
% of extracted variance explained by first-order factors		25.5		68		41.8		21.80		26.10				33.50		30.20		

Notes: GPF, general factors; F1-F3, first-order factors; IB, in-basket exercise; RP, role play exercise; GE, group exercise

Table II.
SLS output and general factors across exercises across three samples

Discussion

The current study aimed to determine the extent of a GPF in AC ratings and to establish if ethnic group differences in ACs were more pronounced on exercises and dimensions that are more analytical by design. The existing literature highlights this feature as a factor resulting in AC performance disparity (Dean et al., 2008; Goldstein et al., 1998, 2001). The study therefore contributes to the literature on a GPF in ACs across exercises and dimensions.

Table III. Black-White means, standard deviations, Cohen's *d* and intercorrelations on AC exercises across three samples

	Sample 1			Sample 2		Sample 3	
	GE	RP	IB	RP	IB	RP	IB
White *M* (SD)	2.46 (0.30)	2.48 (0.26)	2.72 (0.53)	2.42 (0.44)	2.11 (0.50)	2.47 (0.40)	2.10 (0.61)
Black *M* (SD)	2.35 (0.31)	2.33 (0.31)	2.21 (0.55)	2.09 (0.53)	1.77 (0.46)	2.31 (0.47)	1.85 (0.40)
F-value	6.08*	11.63*	37.90*	28.38*	33.81*	30.72*	11.37*
d	0.38	0.53	0.95	0.66	0.73	0.34	0.56
R^2	0.035	0.064	0.182	0.092	0.108	0.026	0.067
1	1	0.280**	0.349**	0.553**	1	0.525**	1
2	0.280**	1	0.219**	1	0.553**	1	0.525**
3	0.349**	0.219**	1				

Notes: GE, group exercise; RP, role play exercise; IB, in-basket exercise. **Correlation is significant at the 0.01 level (two-tailed); *$p < 0.01$

Table IV. Black-White means, standard deviations, Cohen's *d* and intercorrelations on five dimensions for Sample 1

Dimension	White *M* (SD)	Black *M* (SD)	*F*-value	*d*	R^2	1	2	3	4	5
1. BA	2.45 (0.29)	2.15 (0.29)	42.2*	1.00	0.199	1	0.602**	0.527**	0.849**	0.874**
2. C	2.65 (0.25)	2.57 (0.27)	3.49	0.29	0.020	0.602**	1	0.691**	0.660**	0.618**
3. FR	2.63 (0.25)	2.47 (0.28)	14.94*	0.60	0.081	0.527**	0.691**	1	0.633**	0.489**
4. L	2.43 (0.30)	2.16 (0.30)	36.64*	0.93	0.177	0.849**	0.660**	0.633**	1	0.860**
5. RD	2.53 (0.30)	2.25 (0.30)	36.99*	0.94	0.179	0.874**	0.618**	0.489**	0.860**	1

Notes: BA, business acumen; C, communication; FR, fostering relationships; L, leadership; RD, results driven. **Correlation is significant at the 0.01 level (two-tailed); *$p < 0.01$

Table V. Black-White means, standard deviations, Cohen's *d* and intercorrelations on three dimensions for Sample 2

Dimension	White *M* (SD)	Black *M* (SD)	*F*-value	*d*	R^2	1	2	3
1. TM	2.25 (0.42)	1.95 (0.47)	27.65*	0.66	0.090	1	0.744**	0.816**
2. INT	2.36 (0.49)	2.05 (0.55)	22.14*	0.59	0.074	0.744**	1	0.770**
3. PS	2.11 (0.47)	1.75 (0.45)	40.79*	0.80	0.128	0.816**	0.770**	1

Notes: TM, team management; INT, interpersonal interaction; PS, problem analysis and decision making. **Correlation is significant at the 0.01 level (two-tailed); *$p < 0.01$

Table VI. Black-White means, standard deviations, Cohen's *d* and intercorrelations on four dimensions for Sample 3

Dimension	White *M* (SD)	Black *M* (SD)	*F*-value	*d*	R^2	1	2	3	4
1. AST	2.14 (0.52)	1.90 (0.48)	23.79*	0.49	0.053	1	0.770**	0.523**	0.614**
2. PS	2.34 (0.52)	2.12 (0.45)	21.63*	0.47	0.048	0.770**	1	0.580**	0.751**
3. MIC	2.48 (0.50)	2.34 (0.45)	8.95*	0.30	0.021	0.523**	0.580**	1	0.564**
4. PO	2.12 (0.53)	1.87 (043)	26.92*	0.52	0.059	0.614**	0.751**	0.564**	1

Notes: AST, applied strategic thinking; PS, problem analysis and decision making; MIC, managing interpersonal conflict and resolving problems; PO, planning and organizing. **Correlation is significant at the 0.01 level (two-tailed); *$p < 0.01$

A major contribution of the present study is that it is the first, to our knowledge, to examine the extent of a GPF across AC exercises and dimensions in relation to group membership in an

emerging market economy. Furthermore, the study examines multiple samples across private and public organizations in South Africa.

The results in Table II provide support for a GPF in AC exercises across the three samples (H2a). Although previous research has found that GPF accounts for less than 20 percent of the variance in AC ratings (Hoffman et al., 2011), with the exception of the RP exercise in Sample 1 (Table II), GPF accounted for significant variance (approximately 58.2-78.2 percent) in AC ratings in our study. The results reflect similar findings reported in the literature (Lance et al., 2007). The unequal pace of labor market development in South Africa (which may also be true for similar emerging markets) due to access to education, lack of quality education (Ha, 2016), technological skill, and compliance with legislation (e.g. affirmative action policies in South Africa) (Faulk and Salem, 2014) may explain the findings to some extent. We further hypothesize that the SLS might provide more accurate estimates of a GPF as it allows us to identify the proportion of variance attributed to first- and second-order factors.

We were unable to complete the SLS for dimensions in each sample. This may be due to the design of the AC in that the dimensions in the three samples were not purposely designed in a hierarchical manner, a prerequisite to completing the SLS analysis. An unintended outcome of the SLS did allow us to redefine the overarching dimensions for each AC exercise across the three samples, resulting in a more accurate categorization of constructs. Meta-dimensions were therefore identified for each AC exercise after GPF had been taken into consideration. The meta-dimensions were categorized as reflecting three broad categories across the three samples that could be attributed to either task or interpersonal elements (the result of this analysis is available from the authors upon request). This categorization seems consistent with broad dimension categories identified in the literature (Arthur et al., 2003).

The results in Table III provide support for the premise that AC exercises vary in the extent to which they produce group differences (H1a). When reviewing the effect sizes across the three samples, the difference was more pronounced on the IB exercise (Sample 1: d = 0.95; Sample 2: d = 0.73; Sample 3: d = 0.56). This finding is in line with our original hypothesis, showing alignment to previous research (Goldstein et al., 1998; Hoffman et al., 2015), although the magnitude of these differences were larger for this study. The implication of this finding is that the IB exercise is potentially unsuitable for organizations operating in culturally diverse settings and highlights the impact of political, economic, and societal factors at work in these environments.

The results in Tables IV-VI confirmed that dimensions varied in the extent to which they produced group differences (H1b). The results in Sample 1 (Table IV) and Sample 2 (Table V) conformed to our initial hypothesis and showed problem solving dimensions generally produced the largest group differences (Sample 1, BA: d = 1.00; RD: d = 0.94; Sample 2, PS: d = 0.80). However, this finding was not consistent across the three samples. In Sample 1, for example, BA, RD, and leadership (L) yielded the greatest variance in terms of group membership. BA and RD are typically aligned with task or output, while L is focused on people.

In Sample 2, the pattern of findings was confirmed and PS reported the largest group difference. Closer inspection of the dimensions in Table VI revealed that PO (d = 0.52) produced the largest group difference in Sample 3. PO has been found to positively correlate with job performance (Arthur et al., 2003), so this finding is not unusual in the context of the predictive validity of the AC. For the three samples, the reported intercorrelations were moderate to high. This could indicate an overlap of individual items being measured in the dimensions. Although the findings in our study follow the same patterns as the extant literature (Goldstein et al., 2001; Meriac et al., 2014), the magnitude of group differences was more pronounced. We hypothesize that language and education might be factors contributing to larger group differences across exercises and dimensions in emerging markets, such as South Africa.

A novel contribution of the study compared performance in AC ratings between a private organization and a government organization. We expected to (but did not) find a similar pattern of results for Samples 2 and 3 as they were from the same organization in the public sector. Samples 1 and 2 showed greater similarities across dimensions and exercises in relation to Sample 3. In Samples 1 and 2 (both at team leader level), the results for the IB exercise and problem solving dimensions were more closely aligned than the results of the IB and problem solving dimensions for Sample 3 (middle manager level). This finding provides preliminary support for the generalizability of our results in different organizational contexts at the same managerial level. Although the findings in Sample 3 exhibited a similar pattern of results the effect sizes were smaller, showing closer alignment to findings in the international literature (Bobko and Roth, 2013). We hypothesize that job knowledge at higher managerial levels acts as a moderating variable of group differences. Job tenure might therefore reduce group differences due to the acculturation of ethnic groups to the same set of organizational values and behaviors over time.

Furthermore, our study focused on job incumbents being assessed for development purposes. We found that the extent of group differences across dimensions and exercises was closer to the overall ethnic group d's reported for job applicants (Bobko and Roth, 2013). This might suggest that AC research into ethnic group differences in developed countries is not easily generalizable to emerging market economies.

In our study, problem solving dimensions presented as the highest sources of group differences for Sample 1 (BA: d = 1.00; RD: d = 0.94) and Sample 2 (PS: d = 0.80). Even though we did not specifically correlate AC ratings with cognitive criterion, our findings conform to a pattern of results that is consistent with previous AC research showing a relationship between cognitively loaded AC exercises, dimensions, and cognitive ability (Goldstein et al., 1998, 2001; Hoffman et al., 2015; Meriac et al., 2014).

Implications for organizations and employers

This study provides information on the magnitude of a GPF in AC exercises and dimensions, in relation to group membership in an emerging market economy. Our research indicated that the extent of a GPF found in AC ratings was associated with larger group differences. We found evidence to support large group differences across AC exercises and

dimensions that were more cognitively loaded. Together, these factors call into question the practicality of using ACs as an alternative to reduce measuring group differences in assessments (Dean et al., 2008; Goldstein et al., 1998, 2001). Of note is that our study found group differences in AC performance to be larger than those reported in the extant literature (Bobko and Roth, 2013; Hoffman et al., 2015).

The implication is that ACs do not seem to be the "magic bullet" to overcome problems that appear to plague cognitive and trait-based assessments. Our findings show that ACs are equally vulnerable to the perils that persist in these types of assessments when it comes to group differences. Moreover, if Whites and Blacks are equally represented in applicant groups, then cognitively loaded exercises and dimensions will most likely benefit Whites. The most probable explanation is that education and language disparity contributes to AC performance differences across ethnic groups in South Africa.

We offer three recommendations for organizations using ACs in these contexts. First, without compromising the complexity requirements of the intended management level, written exercises should be designed with content that accommodates the language proficiency of different ethnic groups to a reasonable extent. Second, more exercises should be added to the AC and some of these exercises could be designed to reflect specific managerial functions or tasks (Lance, 2008). Third, the over-reliance of cognitively loaded AC exercises and dimensions that are not essential to the job should be avoided.

Limitations and directions for future research

Our study had several limitations that could be addressed in future research. First, even though the ACs were designed according to standard practices and guidelines, not all dimensions were measured across all the exercises. As a result of the SLS procedure only dimensions that were measured multiple times across the exercises were included in the analysis. Future research should aim to include more dimensions and investigate their relation to group differences. Second, only three AC exercises were investigated in this study. The inclusion of additional AC exercises would have strengthened the findings. Future research is required on additional exercises and to replicate the findings of the exercises in our study. Research on both AC exercises and dimensions in other emerging markets is important to determine the generalizability of our findings in similar contexts to South Africa. Third, this study did not consider the impact of assessors on participant performance in the AC. Additional research is needed to investigate this variable in relation to a GPF and group differences. Fourth, we only investigated group differences for Blacks and Whites. This was, in part, due to the comparable sample size of these two subgroups, as well as the vast array of extant literature on the subject. Further research that focuses on other minority groups in South Africa is therefore warranted. Fifth, the study did not include any psychometric tests or criterion-related variables. This type of data would have been useful to rule out cognitive ability as a probable source of group differences in the findings. Sixth, this study made use of a statistical technique not often used in AC research and future research is needed to replicate the findings applying the SLS.

Furthermore, the AC ratings across the three samples were obtained as part of a development exercise. Future research is needed to see if the results will hold in a high stakes selection context. Lastly, additional research is required to determine to effect of moderating variables on group differences across exercises and dimensions, for example, applicants vs incumbents, language, education, and tenure and job knowledge.

Conclusion

The perception that assessments are unfair and discriminatory in their application toward different ethnic groups has been a prevalent concern for practitioners and organizations; especially in emerging market economies, such as South Africa. In light of this it is important to contextualize group differences in multicultural contexts and consider the impact of these differences when using alternative predictors, such as the AC.

Our findings demonstrate that regardless of whether one advocates for dimension or exercise performance in AC ratings, there is a considerable GPF across AC exercises. Significant differences were found between groups in respect of the IB exercise and cognitively loaded dimensions such as BA, problem solving, and strategic thinking. However, we see this evidence as crucial to informing AC design practices to mitigate the influence of group differences. Incorporating the findings of this study in practice will allow for more appropriate design and implementation of ACs in multicultural settings. Our study should encourage future research to investigate additional components of a GPF in AC ratings and in relation to group membership.

Notes

1. The ethnic demographics in South Africa are defined as Black African, White, Colored, Indian/ Asian (Statistics South Africa, 2016). Employment equity legislation defines “Blacks” differently as a generic term that includes Africans, Coloreds, and Indians (Republic of South Africa, 1998). Organizations are, however, required to report on ethnic representation for each separate subgroup for equity purposes according to the SSA definition above. In our study, we use the SSA definition where Blacks refer to Africans and excludes Coloreds and Indians/Asians. We chose to focus on Black-White differences due to the prevalence of research in the extant literature. Additionally, sample sizes for Coloreds and Indians/Asians were too small for reliable estimates.
2. In Sample 1, data were collected over multiple years and were analyzed as one sample. A cross-tab analysis showed there were no significant differences in the spread of ethnicity and gender across the different year groups. A one-way ANOVA was conducted between the year cohorts in Sample 1 on the FDRs. The results showed that there were no significant mean differences at the FDR level across years. We therefore aggregated multiple year data into one sample.

References

Arthur, W.J., Day, E.A., McNelly, T.L. and Edens, P.S. (2003), "A meta-analysis of the criterion-related validity of assessment center dimensions", Personnel Psychology, Vol. 56 No. 1, pp. 125-154.

Bobko, P. and Roth, P.L. (2013), "Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: verifying some perceptions and updating/correcting others", Personnel Psychology, Vol. 66 No. 1, pp. 91-126.

Bowler, M.C. and Woehr, D.J. (2006), "A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings", Journal of Applied Psychology, Vol. 91 No. 5, pp. 1114-1124.

Charoux, J.E. (1997), "Personality testing: be careful!", Management Today, Vol. 13 No. 5, pp. 31-34. Charoux, J.E. and Hurst, D. (1992), "Future potential", People Dynamics, Vol. 10 No. 12, pp. 35-36.

Cohen, J.W. (1988), Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Dean, M.A., Roth, P.L. and Bobko, P. (2008), "Ethnic and gender subgroup differences in assessment center ratings: a meta analysis", Journal of Applied Psychology, Vol. 93 No. 3, pp. 685-691.

Euromonitor Research (2011), "Labour market diversity in emerging market economies", available at: http://blog.euromonitor.com/2011/01/emerging-market-economies-particularly-in-asia-and- latin-america-are-showing-encouraging-signs-of-employment-recovery-from.html (accessed March 25, 2017).

Faulk, G. and Salem, K. (2014), "Emerging opportunities in emerging markets", available at: www.worldwideerc.org/Resources/MOBILITYarticles/Pages/0514Faulk.aspx (accessed March 25, 2017).

Foxcroft, C.D. and Roodt, G. (2005), An Introduction to Psychological Assessment in South Africa, 2nd ed., Oxford University Press, Oxford.

Goldstein, H.W., Yusko, K.P. and Nicolopoulos, V. (2001), "Exploring Black-White subgroup differences of managerial competencies", Personnel Psychology, Vol. 54 No. 4, pp. 783-807.

Goldstein, H.W., Yusko, K.P., Braverman, E.P., Smith, D.B. and Chung, B. (1998), "The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises", Personnel Psychology, Vol. 51 No. 2, pp. 357-373.

Ha, L. (2016), "Rising education standards in emerging market economies will support income and economic growth", available at: http://blog.euromonitor.com/2016/02/rising-education- standards-in-emerging-market-economies-will-support-income-and-economic-growth.html (accessed March 25, 2017).

Hayton, J.C., Allen, D.G. and Scarpello, V. (2004), "Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis", Organizational Research Methods, Vol. 7 No. 2, pp. 191-205.

Hoffman, B.J. (2012), "Exercises, dimensions and the battle of Lilliput: evidence for a mixed-model interpretation of assessment center performance", in Jackson, D.J., Lance, C.E.

and Hoffman, B.J. (Eds), The Psychology of Assessment Centers, Routledge, New York, NY, pp. 281-306.

Hoffman, B.J., Kennedy, C.L., LoPilato, A.C., Monahan, E.L. and Lance, C.E. (2015), "A review of the content, criterion-related, and construct-related validity of assessment center exercises", Journal of Applied Psychology, Vol. 10 No. 4, pp. 1143 – 1168.

Hoffman, B.J., Melchers, K.G., Blair, C.A., Kleinmann, M. and Ladd, R.T. (2011), "Exercises and dimensions are the currency of assessment centers", Personnel Psychology, Vol. 64 No. 2, pp. 351-395.

Jackson, D.J., Stillman, J.A. and Englert, P. (2010), "Task-based assessment centers: empirical support for a systems model", International Journal of Selection and Assessment, Vol. 18 No. 2, pp. 141-154.

Krause, D.E., Kersting, M., Heggestad, E.D. and Thornton, G.C. (2006), "Incremental validity of assessment center ratings over cognitive ability tests: a study at the executive management level", International Journal of Selection and Assessment, Vol. 14 No. 4, pp. 360-371.

Krause, D.E., Rossberger, R.J., Dowdeswell, K., Venter, N. and Joubert, T. (2011), "Assessment center practices in South Africa", International Journal of Selection and Assessment, Vol. 19 No. 3, pp. 262-275.

Kriek, H.J., Hurst, D.N. and Charoux, J.E. (1994), "The assessment centre: testing the fairness hypothesis", Journal of Industrial Psychology, Vol. 20 No. 2, pp. 21-25.

Kuncel, N.R. and Sackett, P.R. (2014), "Resolving the assessment center construct validity problem (as we know it)", Journal of Applied Psychology, Vol. 99 No. 1, pp. 38-47.

Lakens, D. (2013), "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs", Frontiers in Psychology, Vol. 4, 12pp., available at: http:// journal.frontiersin.org/article/10.3389/fpsyg.2013.00863/full doi: 10.3389/fpsyg.2013.00863.

Lance, C.E. (2008), "Why assessment centers do not work the way they are supposed to", Industrial and Organisational Psychology, Vol. 1, pp. 84-97.

Lance, C.E., Foster, M.R., Nemeth, Y.M., Gentry, W.A. and Drollinger, S. (2007), "Extending the nomological network of assessment center construct validity: prediction of cross-situationally consistent and specific aspects of assessment center performance", Human Performance, Vol. 20 No. 4, pp. 345-362.

Lance, C.E., Lambert, T.A., Gewin, A.G., Lievens, F. and Conway, J.M. (2004), "Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings", Journal of Applied Psychology, Vol. 89 No. 2, pp. 377-385.

Meiring, D. (2007), Bias and Equivalence of Psychological Measures in South Africa, Labyrint Publication, Ridderkerk.

Meiring, D. and Buckett, A. (2016), "Best practice guidelines for the use of the assessment centre method in South Africa (5th edition)", SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, Vol. 42 No. 1, 15pp., available at: http://dx.doi.org/10.4102/sajip.v42i1.1298

Meiring, D., Van de Vijver, F.J., Rothmann, S. and Barrick, M.R. (2005), "Construct, item, and method bias of cognitive and personality tests in South Africa", South African Journal of Industrial Psychology, Vol. 31 No. 1, pp. 1-8.

Meiring, D., Van de Vijver, F.J., Rothmann, S. and Sackett, P.R. (2007), "Internal and external bias of cognitive and personality measures in South Africa", in Meiring, D. (Ed.), Bias and Equivalence of Psychological Measures in South Africa, Labyrint Publication, Ridderkerk.

Melchers, K.G. and Annen, H. (2010), "Officer selection for the Swiss armed forces", Swiss Journal of Psychology, Vol. 69 No. 2, pp. 105-115.

Melchers, K.G., Wirz, A. and Kleinmann, M. (2012), "Dimensions and exercises: theoretical background of mixed-model assessment centers", in Jackson, D.J., Lance, C.E. and Hoffman, B.J. (Eds), The Psychology of Assessment Centers, Routledge, New York, NY, pp. 237-254.

Meriac, J.P., Hoffman, B.J. and Woehr, D.J. (2014), "A conceptual and empirical review of the structure of assessment center dimensions", Journal of Management, Vol. 40 No. 5, pp. 1269-1296.

Nell, V. (2000), Cross-Cultural Neuropsychological Assessment: Theory and Practice, Erlbaum, London. Pallant, J. (2013), SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS, McGraw-Hill, Berkshire.

Ployhart, R.E. and Holtz, B.C. (2008), "The diversity-validity dilemma: strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection", Personnel Psychology, Vol. 61, pp. 153-172.

Putka, D.J. and Hoffman, B.J. (2013), "Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings", Journal of Applied Psychology, Vol. 98 No. 1, pp. 114-133.

Republic of South Africa (1998), "Employment equity act (EEA)", Act No. 55 of 1998.

Rushton, J.P., Skuy, M. and Fridjhon, P. (2003), "Performance on Raven's advanced progressive matrices by African, east Indian, and white engineering students in South Africa", Intelligence, Vol. 31 No. 2, pp. 123-137.

Schmidt, F.L. and Hunter, J.E. (1998), "The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings", Psychological Bulletin, Vol. 124 No. 2, pp. 262-274.

Statistics South Africa (2016), "Quarterly labour force survey", available at: www.statssa.gov.za/ publications/P0211/P02111stQuarter2016.pdf (accessed March 10, 2017).

Theron, C. (2007), "Confessions, scapegoats and flying pigs: psychometric testing and the law", SA Journal of Industrial Pyschology, Vol. 33 No. 1, pp. 102-117.

Thornton, G.C., Rupp, D.E. and Hoffman, B.J. (2015), Assessment Center Perspectives for Talent Management Strategies, 2nd ed., Routledge, New York, NY.

Viswesvaran, C., Schmidt, F.L. and Ones, D.S. (2005), "Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences", Journal of Applied Psychology, Vol. 90 No. 1, pp. 108-131.

Wolff, H. and Preising, K. (2005), “Exploring item and higher order factor structure with the Schmid- Leiman solution: syntax codes for SPSS and SAS”, Behavior Research Methods, Vol. 37 No. 1, pp. 48-58.