European Journal of Personality
Eur. J. Pers. **18**: 127–141 (2004)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/per.504

Three Method Factors Explaining the Low Correlations between Assessment Center Dimension Ratings and Scores on Personality Inventories†

NANJA J. KOLK[1]*, MARISE PH. BORN[2] and HENK VAN DER FLIER[3]

[1]*The Change Factory, Berenschot Business Consultants, Utrecht, The Netherlands*
[2]*Department of Social Sciences, Erasmus University, Rotterdam, The Netherlands*
[3]*Department of Work and Organizational Psychology, Vrije Universiteit Amsterdam, The Netherlands*

Abstract

In general, correlations between assessment centre (AC) ratings and personality inventories are low. In this paper, we examine three method factors that may be responsible for these low correlations: differences in (i) rating source (other versus self), (ii) rating domain (general versus specific), and (iii) rating format (multi- versus single item). This study tests whether these three factors diminish correlations between AC exercise ratings and external indicators of similar dimensions. Ratings of personality and performance were combined in an analytical framework following a $2 \times 2 \times 2$ *(source, domain, format) completely crossed, within subjects design. Results showed partial support for the influence of each of the three method factors. Implications for future research are discussed. Copyright © 2004 John Wiley & Sons, Ltd.*

It is well known that assessment centres (ACs) measure job relevant constructs, thanks to a satisfactory criterion-related validity (Gaugler, Rosenthal, Thornton, & Bentson, 1987). Unfortunately, we do not know exactly what these constructs are (Russell & Domm, 1995). This question has resulted in an abundant body of research seeking the meaning of the AC dimensions. These studies have mainly focused on the effects of the AC architecture (dimensions, exercises, assessor characteristics, etc) on construct validity (see Lievens and Conway, 2001, for a review). Evaluating the results of two decades of research, Lance et al. (2000, p. 344) noted that 'In retrospect, we think the question ''we know what the assessment center constructs are (i.e., dimensions); are these valid?'' was

*Correspondence to: Nanja J. Kolk, Berenschot, Europalaan 40, 3526 KS Utrecht, The Netherlands.
E-mail: n.kolk@berenschot.com
†An earlier version of this paper was presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, April 2001.

Contract/grant sponsor: Laboratory for Applied Psychology (LTP), Amsterdam.

Received 2 December 2002
Accepted 29 July 2003

premature'. Lance et al. proposed that Russell's (1994) question, 'what are the assessment center constructs?', needed to be answered first. To answer the latter question, construct-related validity of ACs has been studied by placing AC dimension ratings in a nomological network of cognitive ability tests and personality inventories.

While some of the results using the external construct validity approach were actually promising (Scholz & Schuler, 1993; Shore, Shore, & Thornton, 1992; Shore, Thornton, & Shore, 1990), other studies failed to show the expected relationships between AC dimension ratings and these external measures. Specifically, it appeared that cognitive ability showed some relatedness with AC ratings, while the AC–personality inventory correlation was often negligible (see e.g. Borman, 1982; Bray & Grant, 1966; Chan, 1996; Crawley, Pinder, & Herriot, 1990; Fleenor, 1996; Hinrichs, 1978; Lance et al., 2000; Tziner & Dolan, 1982). This latter finding appears to be yet another in a long line of evidence against the AC dimension's construct validity, adding concerns about external validity to a larger literature on internal construct validation problems (see e.g. Chan, 1996; Lievens & Conway, 2001; Sackett & Dreher, 1982). For instance, Crawley et al. (1990, p. 215) note that 'The general implication for the assessment centre design is to cast further doubt on the use of the ''sign'' as opposed to the ''sample'' rationale'. Crawley et al. propose that the exercises—based on job analysis—are the units of measurement, thereby abandoning the use of dimensions entirely. Recently, Lance et al. (2000) called for additional research using exercise factors (thereby abandoning the dimensions) and relating these factors to external measures thought to be associated with AC performance. However, in our opinion, ACs and personality inventories—which are often used as these external measures—differ on three method factors that may provide an explanation for the weak relationships. Therefore, before turning to the nomological network approach of studying relationships between AC dimensional ratings and external measures, the appropriateness of this validation strategy must be warranted. This study examines whether the correlations between scores on personality inventories and scores on AC dimensions are affected by differences in (i) rating source, (ii) rating domain, and (iii) rating format.

First, AC performance is rated by assessors ('others') and a personality inventory is rated by applicants ('self'). It is well documented that correlations between self- and other-ratings (peers, supervisors, subordinates, etc.) are significantly lower than correlations between several other-ratings (see e.g. Furnham & Stringfield, 1998; Harris & Schaubroeck, 1988; Kenny, 1994; Murphy & Cleveland, 1995). This consistent and pervasive finding is called self–other rating disagreement (see e.g. Cheung, 1999). Thus, the AC–personality inventory correlation is hypothesized to be diminished by self–other rating disagreement.

Second, ACs and personality inventories differ in measurement domain. A study by Goffin, Rothstein, and Johnson (1996) showed that personality inventories have incremental validity over that of the AC in predicting performance. This result led these researchers to conclude that '. . . personality and AC scores may sample different domains, which in turn may predict relatively independent aspects of the domain of performance' (Goffin et al., 1996, p. 753). For this reason, low AC–personality inventory correlations are not problematic, according to Goffin et al. (1996), since the two predictors seem to measure different things. Specifically, general performance measures can obscure potentially important distinctions in how targeted traits may be related to specific work behaviours (Tett, Guterman, Bleier, & Murphy, 2000, p. 213). Within the AC context, this implies that domain specificity may limit the AC–personality inventory correlation. That

is, the AC dimension is domain specific, whereas the personality measure is not linked to a specific managerial behaviour domain. Thus, rating domain specificity is the second method factor potentially diminishing AC–personality inventory relationships.

A third method factor possibly diminishing the AC–personality inventory correlation is rating format. Although Harris and Schaubroeck (1988) have shown rating format not to moderate the self–other rating correlation, format influences on the AC–personality inventory correlation have not been examined. Moreover, Schneider and Schmitt (1992) have shown exercise format (individual versus group exercises) to affect the relationships among exercise ratings. It is to be expected that the influence of format of the stimulus materials (exercises, rating forms, etc.) is also present when exercises and personality inventories are considered, because these methods have an even more dissimilar format than two AC exercises.

To substantiate the influence of the factors rating source, format, and domain on the AC–personality inventory relationship, this study considers correlations between AC ratings and measures varying on these three factors. We have configured the three factors as follows. Rater type is measured by self versus other (assessor or peer rating). Rating domain is measured by a general (personality inventory, nonspecific rating scale) versus a specific (managerial situations) measuring device. Rating format is measured by an AC exercise format (single-item scale) versus a questionnaire format (multi-item scale). The three factors are studied simultaneously yet independently in a $2 \times 2 \times 2$ fully crossed within subjects analytical framework (Table 1).

Hence, each of the three factors was measured by two poles: self versus other rating source; general versus specific rating domain; multi-item versus single-item rating format. We hypothesize that each of the three factors influences the AC–personality inventory correlation significantly. More specifically, we hypothesize the following relationships.

(i) Regarding the influence of *rating source* on the AC–personality inventory correlation, we hypothesize that the correlation between AC other and personality inventory other ratings (both other rating), as well as the correlation between AC self and personality inventory self rating (both self rating) exceeds the AC other–personality inventory self correlation.

(ii) Regarding the influence of *measurement domain*, we expect that a situation response inventory (SRI) measuring the same level of situational specificity as the AC relates more to the AC than a general personality inventory. Also, two measurements of a

Table 1. $2 \times 2 \times 2$ (ratings source, domain, format) design

	Rating source			
	Self		Other	
	Rating format			
	Multi-item	Single item	Multi-item	Single item
Rating domain				
General domain	NEO self	General rating self	NEO other	General rating other
Situation specific domain	SRI self	AC self	SRI other	AC other

AC, assessment centre exercise; SRI, situation response inventory.

general domain should relate more to one another than the basic AC other–personality inventory self correlation. Hence, the second hypothesis predicts that AC other–SRI self ratings and general other ratings–personality inventory self ratings correlate higher than AC other–personality inventory self rating.

(iii) A third set of hypotheses concerns the factor *rating format*. The AC exercise rating involves a direct measurement of the construct on a single-item rating scale, whereas personality is judged indirectly on a multi-item scale. If format affects the AC other–personality inventory self correlation, AC other ratings and general self ratings (both single-item scales), as well as personality inventory self and SRI other ratings (both multi-item scales) should correlate higher than the basic AC other–personality inventory self correlations.

METHOD

Procedure

Participants' performance in this study was rated though multiple measures by multiple raters: self and other ratings of an AC exercise, personality inventories, and general rating scales. These data were collected in seven 1 day developmental ACs. During each AC, participants were asked to take part in an exercise and fill out tests and inventories. At the end of the day, they were handed several inventories and were asked to give these inventories to one or two peers to fill out. They could choose whomever they thought most capable of making inferences about their behaviour.

Ratings

Participants

Participants were 149 men and women (37 per cent male), a few months before or after reaching their master's degree (mostly in economics, psychology, or law). Their mean age was 26 years (SD = 5). The incentive for participation in the study was training and multi-source feedback on their AC performance, so as to prepare themselves for coming job interviews and ACs.

Assessors

Participants were rated by a pool of 23 professional assessors and 11 role-players. Both the assessors and the role-players received recurring assessor training sessions, focusing on the meaning of the dimensions, on rating errors, and on a shared frame of reference. The exercise ratings (see description under 'Measure') were made on Likert-type interval scales, ranging from 1 (low) to 5 (high). Inter-rater reliabilities were moderate (i.e. the mean PPM correlation coefficient $r = 0.63$). The rater–ratee ratio was two raters (the assessor and the role-player) to one ratee (the participant).

Peers

Each participant was also assessed by one or two peers (depending on the peers' response rate). These peers were family members, friends, or colleagues of the participants. They volunteered to participate in order to provide multi-source feedback to the participants.

Because there is only one 'self', for comparison reasons we used only one of the two exercise ratings (i.e. the rating by the assessor) and one of the two peer ratings (at random), such that each rating was made by only one person per rating source: the participant him/herself, one assessor, and one peer.

Measures

As a representative for the AC, we used an *interview simulation exercise*, in which participants were asked to persuade a subordinate to comply with a certain task. This type of exercise appeared to have the highest or second highest correlation with overall assessment rating (OAR) compared with other exercises (Thornton & Byham, 1982) and may be regarded as a fairly typical exercise, as it is used in 47 per cent of all ACs (Thornton, 1992). The exercise took 15 min for preparation and 15 min to play. Participants rated themselves similarly, after becoming acquainted with the meaning of the dimensions. They were assured that these self-ratings would not affect their overall assessment performance ratings. To ensure this study's fidelity, all AC materials (dimension definitions, rating forms, and exercise instructions) were obtained from a working AC.

Two personality inventories were used in the present study. For the general personality inventory we administered the NEO Personality Inventory—Revised (Costa & McCrae, translation by Hoekstra, Ormel, & De Fruyt, 1996). The NEO is a self-report personality test, which consists of 240 items, measuring the Big Five domains (Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness) on five-point scales (ranging from 1, 'strongly disagree', to 5, 'strongly agree'). Each domain consists of six facets. Research supports the psychometric qualities (validity and reliability) of the NEO (McCrae & Costa, 1989, 1992). Peers also rated the participants on the NEO, which was redesigned for the present study (with permission from the authors: Hoekstra et al., 1996). Specifically, 'I like . . . ' was replaced by names of the target assessees, so 'Jim likes . . . ' was the wording when the participant's name was Jim.

The second questionnaire administered to the participants and their peers was a situation response inventory (SRI). The purpose of the SRI was to measure the same constructs as the AC, yet in a questionnaire format. Therefore, we could not rely on existing materials, but needed to develop a new inventory especially for the present study (Kolk & van den Acker, 1998). The SRI consisted of multiple sets of situations and responses. Each set started with a description of an interpersonal management situation, followed by six to eight possible reactions to this situation. The situations were derived from conventional AC role-plays: interpersonal management situations (e.g. difficult discussion with an unwilling subordinate). The responses were based on typical dimension-related responses of candidates in an AC role play (e.g. describing attempts to act sensitively). Six subject matter experts (HR professionals with more than 3 years of experience in rating candidates in AC role-plays) evaluated the content validity of the SRI. First, they rated on a five-point scale whether the descriptions of the situations were representative of AC role-plays. We then discarded all situations with an average score less than 4.0. Second, we asked the experts to sort the responses into the dimension-categories sensitivity, analytical skills, persuasiveness, or more of these. We rejected all responses that were not appropriately categorized (i.e. hit rate <80 per cent). The final situation response inventory consists of 15 management situations and 43 responses. Alpha reliabilities for self-ratings were $\alpha = 0.83$, 0.64, and 0.73. The reliabilities of peer ratings were respectively $\alpha = 0.86$, 0.79, and 0.69.

In the SRI, participants and peers were asked to rate on five-point scales whether they (i.e. the target assessees) would react accordingly in these situations. Participants and peers also made an assessment of the participants' behaviour in general (i.e. outside the AC). These *'general' ratings* were made on rank order scales, as opposed to directly on a five-point scale, to decrease social desirability. So, participants and peers were asked to rank order the three dimensions (see 'Target dimensions') regarding the participant's behaviour in general, yielding a three-point measurement scale.

Target dimensions

The targeted dimensions in this study were Sensitivity (creating a pleasant atmosphere during a meeting, being friendly and understanding to others, showing interest in the feelings and needs of others, being a good listener, etc), Analytical Skills (identifying problems, searching for additional information, distinguishing between matters of primary and secondary importance, developing courses of action, having insight to the heart of the matter, etc), and Persuasiveness (speaking firmly and self-confidently, persuading others of a certain viewpoint, staying with the initially held position, handling pressure and stress situations well, etc). These dimensions are quite typical for AC practices (Thornton & Byham, 1982). Also, a previous study, examining AC ratings of more than 1500 job applicants, confirmed that these dimensions were relatively independent (Kolk, Born, & Van der Flier, in press).

AC dimension scores are often aggregated in an OAR, which in turn is correlated with external measures such as general intelligence and personality (see e.g. Scholz & Schuler, 1993). In doing so, a construct (e.g. 'g', conscientiousness) is in effect compared with a method (the AC), which has been postulated to be neither scientifically nor conceptually informative (Arthur, McNelly, Edens, & Day, 2001). Therefore, it is important to seek conceptual similarities between external constructs (i.e. personality scales and facets) and the AC constructs (i.e. dimensions), rather than using the aggregated OAR.

As there was no empirical body of literature to refer to in this respect, the authors of this study independently inferred, on rational grounds, which target dimensions matched the NEO-PI personality traits, by looking for conceptual similarities between those dimensions and the NEO scales and facets. While in some instances the match between an AC dimension and the NEO facets could be theoretically endorsed, other relationships were less obvious. In those cases we selected facets that bore the greatest possible resemblance to the description of a dimension.

First, it was assumed that the Big Five factor Agreeableness was related to the target dimension Sensitivity. Second, Conscientiousness was considered to be related to Analytical Skills. (In this respect, we did not include facets from Openness to Experience, because in the NEO-PI this scale is not operationalized as a cognitive/analytical trait. Rather, it consists of facets referring to active imagination, appreciation for the aesthetic, attention to one's own inner feelings, intellectual curiosity, openness to values, and preference for variety.) Third, we posed that some facets of both Extraversion and Neuroticism were conceptually related to Persuasiveness. This resulted in the following facet compositions: Sensitivity was composed by the Agreeableness facets Trust (A1), Straightforwardness (A2), Altruism (A3), Compliance (A4), Modesty (A5), and Tender-mindedness (A6). Analytical Skills was composed by the Conscientiousness facets Competence (C1), Order (C2), Dutifulness (C3), Achievement striving (C4), Self-discipline (C5), and Deliberation (C6). Persuasiveness was composed by the Extraversion

facets Assertiveness (E3) and Activity (E4) and the (reversed) Neuroticism facets Self-consciousness (N4) and Vulnerability (N6).

Analyses

There was no omnibus test for examining the analytical framework used in this study. Therefore, we analysed correlations between the eight cells in the $2 \times 2 \times 2$ design step by step, through examination of the correlations between each pair of cells that were hypothesized to be stronger than the 'regular' correlation between AC assessor ratings and NEO self ratings, i.e. the *baseline correlations*. Thus, the baseline correlations (AC other–NEO self) were compared with the correlations between measures, which each avoided one of the three hypothesized method factors (see Table 1).

The results were studied per dimension, because the rank order data we used for the general self and general other rating (see Table 1) did not allow for a simultaneous examination of the dimensions, since they were interdependent. Another reason to examine differences on a dimension level was that previous research on 360° ratings has shown that the mere nature of a dimension (i.e. their observability) influences the congruence between rating sources: less observable cognitive dimensions have been shown to lead to less congruence between rating sources than overt dimensions, such as forward planning and communications (Furnham & Stringfield, 1998).

RESULTS

Before turning to the composite NEO scales described above, we report the correlations between the AC assessor ratings on the three dimensions and the regular NEO self ratings. As we expected the correlations were low. Of the 105 correlations (five scales and 30 facets correlated with three dimensions), only three were significant (between Sensitivity and the Fantasy facet of Openness to Experience: $r = 0.23$, $p < 0.001$; between Sensitivity and the Ideas facet of Openness to Experience: $r = 0.19$, $p < 0.05$; and between Sensitivity and the Altruism facet of Agreeableness: $r = 0.19$, $p < 0.05$).

Subsequently, we examined the correlations between self and other ratings on the AC exercise, the general rating, the NEO PI R, and the SRI. Descriptive statistics (means, standard deviations, and correlations) for the 24 scales in the $2 \times 2 \times 2$ design (eight measures and three dimensions) are shown in the Appendix.

Table 2 shows an overview of the correlations between the measures that were hypothesized to exceed the baseline through avoiding each of the three method factors. The first row of correlations in Table 2 represents the baseline of AC other–NEO self ratings. The rating source factor was avoided in correlations 1(a) and (b). The difference in rating domain was avoided in correlations 2(a) and (b). The rating format factor was avoided in correlations 3(a) and (b).

Table 2 shows first that the mean baseline correlation between AC other ratings and NEO self ratings was $r = 0.05$.

Rating source

The first hypothesis predicted that correlations between AC self and NEO self, as well as correlations between AC other and NEO other, would exceed the baseline AC other–NEO self correlation.

Table 2. Summary of the correlations between the ratings on the simulation exercise, the general rating, the NEO PI R, and the situation response inventory

	Sensitivity	Analytical skills	Persuasiveness	Mean
Baseline: AC other–NEO self ($N = 149$)	0.12	0.00	0.04	0.05
1(a) AC self–NEO self ($N = 99$)	0.03	0.24*	0.14	0.14
1(b) AC other–NEO other ($N = 50$)	0.16	0.10	0.26*	0.17
2(a) AC other–SRI self ($N = 145$)	0.10	0.09	0.04	0.08
2(b) GR other–NEO self ($N = 117$)	0.37**	0.19*	0.28**	0.28
3(a) GR self–AC other ($N = 111$)	0.14	0.13	0.03	0.10
3(b) SRI other–NEO self ($N = 121$)	0.34**	0.17*	0.27**	0.26

AC, assessment centre exercise; GR, general rating; NEO, NEO PI; SRI, situation response inventory.
1(a), (b) rating-source method factor avoided.
2(a), (b) rating-domain method factor avoided.
3(a), (b) rating-format method factor avoided.
**$p < 0.01$; *$p < 0.05$, one sided.

The mean correlation 1(a) in Table 2 was $r = 0.14$, which is higher than the mean baseline correlation of $r = 0.05$. Mainly, the Analytical Skills correlation of $r = 0.24$ accounted for this result, being significantly higher than the baseline of $r = 0.00$ ($z = 1.85$, $p = 0.03$).

Worth mentioning is that self AC ratings showed less variance than assessor AC ratings (the mean standard deviations were 0.91 and 0.99 respectively). Self ratings showed a mean of $M = 3.5$ versus an assessor rating mean of $M = 2.6$. On all three dimensions, self ratings were significantly higher than assessor ratings (Analytical Skills, mean difference $d = 0.82$, $t[98] = 6.68$, $p = 0.00$; Persuasiveness, mean difference $d = 0.69$, $t[97] = 6.01$, $p = 0.00$; Sensitivity, mean difference $d = 0.96$, $t[98] = 8.64$, $p = 0.00$). These results, showing that self ratings are more lenient and less variable, are consistent with previous research examining self versus supervisor (Cheung, 1999) or self versus assessor ratings (Clapham, 1998).

The second part of the first hypothesis concerned AC other ratings (assessors) and NEO other ratings (peers). The mean correlation 1(b) was $r = 0.17$, which is higher than the baseline of $r = 0.05$. The Persuasiveness AC other–NEO other correlation of $r = 0.26$ was significantly higher than the baseline of $r = 0.04$ ($z = 2.03$, $p = 0.02$).

Examination of the correlation coefficients 1(a) and (b) (mean $r = 0.15$) calls for the notion that the rating source indeed seems to affect the basic AC other–NEO self correlation, albeit that not all correlations differ significantly from the baseline.

Rating domain

The second hypothesis concerning the influence of rating domain, predicted that AC other–SRI self ratings and general other–NEO self ratings would correlate higher than the baseline AC other–NEO self correlation.

Row 2(a) in Table 2 shows the correlations between SRI self and AC other ratings, which are two situation specific domains. Dimension correlations were not significantly different from the baseline. The mean correlation was $r = 0.08$, indicating a minimal and non-significant difference from the baseline of $r = 0.05$.

Regarding the correlation between two general domains, i.e. rank other ratings and NEO self ratings, row 2(b) in Table 2 shows a mean Spearman rank correlation of $r = 0.28$. This

value is clearly much higher than the basic AC other–NEO self correlation. For comparison reasons, we also looked at Pearson correlations, which appeared to be quite similar. These Pearson correlations were used in the analysis. Results revealed that, on all three dimensions, the correlations between general rank other rating and NEO self rating were significantly higher than the baseline correlations (respectively $z = 1.83$, $p = 0.03$; $z = 1.45$, $p = 0.07$; $z = 1.66$, $p = 0.05$).

Correlations 2(a) and (b), which avoided the rating domain factor, showed mixed support for the hypothesis (overall mean $r = 0.18$); two general measures correlated highly, but two specific measures did not correlate substantially higher than the baseline.

Rating format

Regarding rating format, the third hypothesis stated that AC other ratings–general self ratings and SRI other ratings–NEO self ratings would correlate stronger than the basic AC other–NEO self correlation.

Row 3(a) shows a Spearman rank correlation between general self ratings and assessor dimension ratings. The mean correlation was $r = 0.10$, which is higher than the baseline of $r = 0.05$, yet not significantly higher.

The correlations between NEO self ratings and SRI other ratings are shown in row 3(b). The correlation patterns confirmed what was expected, Agreeableness correlating significantly with Sensitivity, Conscientiousness with Analytical Skills, and Extraversion with facets of Persuasiveness and Neuroticism. The mean correlation was $r = 0.26$, which exceeded the basic AC other–NEO self correlation of $r = 0.05$. The differences between these correlations and the baseline correlations were all significant at the 0.05 or 0.10 level (respectively $z = 1.91$, $p = 0.03$; $z = 1.34$, $p = 0.09$; $z = 1.62$, $p = 0.05$).

Regarding the influence of rating format, results revealed two correlations (mean $r = 0.18$) that exceeded the basic correlation. Yet, two multi-item scales intercorrelated higher than two single-item scales.

DISCUSSION

This study sought evidence that three method factors diminish the relation between assessment centre (AC) exercise ratings and personality inventories, namely rating source, rating domain, and rating format. If this were to be established, the often reported low correlation between assessor ratings and self ratings on personality inventories could be explained. The results of this study lend partial support to the notion that differences in rating source, domain, and format have to be taken into account in drawing inferences from evidence for the ACs external construct validity, using personality inventories as the external measures. All but two of 18 correlations that accounted for these factors were higher than the basic AC other–NEO self correlation. Yet, not all differences were significant, and the correlations remained relatively low, even when the method factors were removed.

In addition, results did not reveal apparent differences in rater congruence between more or less observable dimensions, as was found in the Furnham–Stringfield (1998) study using a 360° feedback inventory. Specifically, the least observable dimension—Analytical Skills—was no less agreed upon by the multiple rating sources than the more observable dimensions of Persuasiveness and Sensitivity.

Regarding the *rating source factor*, the results are in line with earlier findings from performance appraisal and multi-source feedback literature, indicating that other–other agreement exceeds self–other agreement (see e.g. Atwater, Ostroff, Yammarino, & Fleenor, 1998; Cheung, 1999; Kenny, 1994; Murphy & Cleveland, 1995). Therefore, the commonly reported self–other rating disagreement seems to apply in AC settings as well as in other performance appraisal conditions. Multiple other rating sources are somewhat more in agreement than self–other rating sources, as are multiple self-ratings. In the extensive body of literature on this subject, multiple reasons have been offered for self–other rating disagreement. For instance, participants can be expected to use information not made available in the AC exercises, whereas assessors are always constrained to this particular source of information. Also, self-ratings may rely on pre-existing self-schemas, thoughts, and feelings, which may be relevant, but which are obviously not available to assessors (Clapham, 1998; Shore, Tetrick, & Shore, 1998).

As for the influence of *rating domain factor*, the analyses revealed mixed results. The first part of this hypothesis, which stated that two similar specific domains (the situation response inventory versus the AC exercise) would correlate were highly than a general versus a specific domain (NEO versus AC), could not be confirmed. The correlations were not significantly higher than the baseline. However, the second part, which stated that two general domains (NEO versus general rating) would correlate significantly more strongly, was supported by the data.

The third hypothesis predicted that the *rating format factor* would decrease the AC other–NEO self correlation. Indeed, two multi-item scales (the NEO and the situation response inventory) correlated significantly more highly than the baseline AC other–NEO self correlations. As to the general self–AC other correlation, results show that, although the correlations were higher, they were not significantly different from the basic AC other–NEO self correlation. An alternative explanation is that because we used rank-order data for the 'general' self-ratings, the rating scale changed compared with the Likert-type scale in the exercise rating. General ratings could have been made on similar five-point Likert scales. Nonetheless, the pervasive self-enhancement and overestimation tendency of self- and peer raters may cause decreased variance in Likert ratings and was therefore not used.

Seemingly, two general measures correlate more highly than two situation specific measures. Also, it seems that two multi-item scales correlate more highly than two single-item scales. An alternative explanation for these results is offered by the fact that the rating domain and format measures consisted of self- versus assessor ratings, as well as self- versus peer ratings. Although assessors and peers are both 'others', it is conceivable that they evaluate the participant in different ways, particularly because in our study peers were not present during the AC. Support for this contention was offered in a study by Shore et al. (1998), which revealed that peers and assessors weighted the participant's performance information quite differently.

A recent study by Cheung (1999) disentangled several forms of disagreement between self- and other ratings, and categorized these forms into conceptual disagreement (i.e. how the construct is perceived by the rating sources) and psychometric disagreement (i.e. the psychometric properties of the scales used by the ratings sources). The methods described in the study by Cheung focused primarily on self–other rating disagreement, yet generalize to any form of disagreement between two sources (Cheung, 1999, p. 2). Although testing for these different forms of disagreement goes beyond the scope and aim of this paper, the theoretical principles may explain the difference between assessors and peers, which was found in this study, as well as in the Shore et al. (1998) study.

First, peers and assessors are likely to have utilized *different frames of reference* in their assessments of the participants' performance. More specifically, they may have disagreed on the relationship between specific behaviours and underlying performance dimensions. This is even more likely considering that the specific behaviours available to the rating sources were different for groups of raters, since there is different information provided to peers and assessors. Assessors had performance information from participants in a role-playing exercise, which was not accessible to peers. Conversely, peers had information sources not available to assessors, due to their (informal) acquaintance with the participant. In other words, the assessors and peers may have used different information grounds on which they based their assessment (see e.g. Shore et al., 1998, p. 97, 1992).

Cheung (1999) also discussed *source-specific biases*, which lead to rater disagreement (psychometric disagreement); peers and assessors may be biased in different directions. Indeed, both self- and peer ratings are commonly noted to be affected by over-estimation and social desirability bias, producing lenient ratings (Borkenau & Ostendorf, 1989; Hofstee, 1994; Shore, Shore, & Thornton, 1992). Assessors, on the other hand, do not have this leniency tendency, and may even be more likely to rate severely (cf. Cheung, 1999). Results of this study seem to support this hypothesis. This could increase the self–peer correlation compared with the self–assessor and the peer–assessor correlation.

Study limitations

Constraints of this study include the limited sample size and missing data in some of the measurements. These factors may have hindered reaching a level of significance for some of the smaller differences between the correlations.

A second limitation of this study pertains to the use of peers and assessors as 'others'. The results might have been more in line with the hypotheses if we had been able to use a more homogeneous group of 'others'. The same objection may be raised to the use of rank-order scales versus Likert-type scales.

A third limitation is that this study included only one exercise, whereas a working AC usually involves more than one exercise. However, the addition of exercises would make this study even more complex (due to the fully crossed design).

A fourth limitation was indicated by an anonymous reviewer. It concerns the notion that the SRI differed not only in terms of rating domain (personality versus AC constructs) but also in terms of situational specificity. Inherent in the process of designing a questionnaire that aims to measure the same constructs as the AC is the inclusion of managerial situations that prompt the AC constructs (in the absence of any knowledge about possible underlying psychological traits). Thus, a side-effect of creating an AC-based questionnaire is indeed the introduction of situational specificity.

Conclusion

In sum, the present study reveals partial support for the three factors diminishing AC other–NEO self correlations. Rating source, domain and format turned out to have effects on the AC other–NEO self relationship. All mean correlations increased when the influence of each of the three factors was avoided, although the influence of each factor was not always unequivocal. Conclusions should therefore be drawn cautiously. Nevertheless, having ascertained that the three factors at least show some effect on the AC–personality inventory relation, we can conclude that self-report personality inventories have limitations as external validation measures. Rather, we suggest that inquiries into the

relationship between AC exercise ratings and other measuring devices such as personality inventories and 360° feedback be used as bits of information in answering the important question raised by Russell (1994): 'What *does* the assessment center measure?'. Thus, we see eye to eye with Lance et al. (2000) in their call for more inquiries on the AC's validity using the nomological network approach in order to gain a more thorough understanding of what the AC constructs are. Yet, the results of these inquiries within the AC's nomological net should be interpreted while taking the possible influence of method factors into account.

ACKNOWLEDGEMENTS

This research was funded by the Laboratory for Applied Psychology (LTP), Amsterdam. We acknowledge Peter Dekker for his helpful comments on the analyses. We also thank Robert G. Jones for his helpful comments on an earlier version of this manuscript.

REFERENCES

Arthur, W., Jr., McNelly, T. L., Edens, P. S., & Day, E. A. (2001, April). *Distinguishing between methods and constructs: The criterion-related validity of assessment center dimensions*. Paper presentated at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self–other agreement: Does it really matter? *Personnel Psychology, 51*, 577–598.

Borkenau, P., & Ostendorf, F. (1989). Descriptive consistency and social desirability in self- and peer reports. *European Journal of Personality, 3*, 31–45.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67*, 3–9.

Bray, D. W., & Grant, D. L. (1966). The assessment center in the measurements of potential for business management. *Psychological Monographs, 80*, whole No. 625.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69*, 167–181.

Cheung, G. W. (1999). Multifaceted conceptions of self–other ratings disagreement. *Personnel Psychology, 52*, 1–36.

Clapham, M. M. (1998). A comparison of assessor and self dimension ratings in an advanced management assessment center. *Journal of Occupational and Organizational Psychology, 71*, 193–203.

Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology, 63*, 211–216.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319–335.

Furnham, A., & Stringfield, P. (1998). Congruence in job-performance ratings: A study of 360″ feedback examining self, manager, peers, and consultant ratings. *Human Relations, 51*, 517–530.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Goffin, R. D., Rothstein, M. G., & Johnston, N. G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81*, 746–756.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self–supervisor, self–peer, and peer–supervisor ratings. *Personnel Psychology, 41*, 43–62.

Hinrichs, J. R. (1978). An 8-year follow-up of a management assessment center. *Journal of Applied Psychology, 63*, 596–601.

Hoekstra, H. A., Ormel, J., & De Fruyt, F. (1996). *NEO PI-R/NEO FFI; Big Five persoonlijkheidsvragenlijsten* [*NEO PI-R/NEO FFI; Big Five personality inventories*]. Lisse: Swets & Zeitlinger.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8*, 149–162.

Kenny, D. A. (1994). *Interpersonal perception, a social relations analysis*. New York: Guilford.

Kolk, N. J., Born, M. P., & Van der Flier, H. (in press). A triadic approach to assessment center's construct validity: The effect of categorizing dimensions into a feeling, thinking, power taxonomy. *European Journal of Psychological Assessment*.

Kolk, N. J., & van den Acker, C. J. M. (1998). *Leadership Situations Inventory.* Amsterdam: LTP [Laboratory for Applied Psychology].

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. B. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323–353.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait–multimethod studies. *Journal of Applied Psychology, 86*, 1202–1222.

McCrae, R., & Costa, P. (1989). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81–90.

McCrae, R., & Costa, P. (1992). Discriminative validity of the NEO-PIR facet-scales. *Education and Psychologal Measurement, 52*, 229–237.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Russell, C. J. (1994). A model of assessment center construct space and an agenda for future research. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Nashville, TN.

Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology, 68*, 25–47.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401–410.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*, 32–41.

Scholz, G., & Schuler, H. (1993). The nomological network of the assessment center: A metaanalysis. *Zeitschrift für Arbeits-und Organisationspsychologie, 37*, 73–85.

Shore, L. M., Tetrick, L. E., & Shore, T. H. (1998). A comparison of self-, peer, and assessor evaluations of managerial potential. *Journal of Social Behavior and Personality, 13*, 85–101.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42–54.

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101–116.

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a 'hyperdimensional' taxonomy of managerial competence. *Human Performance, 13*, 205–251.

Thornton, G. C. (1992). *Assessment centers in Human Resource Management*. Reading, MA: Addison-Wesley.

Thornton, G. C. I., & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York: Academic.

Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67*, 728–736.

APPENDIX: DESCRIPTIVE STATISTICS OF THE RATINGS ON THE SIMULATION EXERCISE, THE GENERAL RATING, THE NEO PI R, AND THE SITUATION RESPONSE INVENTORY

	M	SD	*n*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Other AC sensitivity	2.57	1.01?	149	—																						
2. Other AC anal. skill	2.59	0.96	149	0.38*	—																					
3. Other AC persuasiven	2.73	1.00	149	0.19*	0.56*	—																				
4. Self AC sensitivity	3.57	0.90	99	0.34*	0.25*	0.20*	—																			
5. Self AC anal skill	3.36	0.88	99	0.11	0.09	0.06	0.57*	—																		
6. Self AC persuasiven.	3.46	0.96	98	0.11	0.32*	0.34*	0.59*	0.59*	—																	
7. Self GR sensitivity	2.21	0.84	111	0.14	−0.12	−0.16	0.07	−0.01	−0.07	—																
8. Self GR anal skill	1.99	0.84	111	0.00	0.14	0.12	0.00	0.02	−0.02	−0.63*	—															
9. Self GR persuasiven	1.80	0.72	111	−0.17	−0.02	0.04	−0.08	−0.01	0.10	−0.44*	−0.42*	—														
10. Other GR sensitivity	2.03	0.89	117	0.11	−0.12	−0.11	−0.00	−0.01	−0.07	0.43*	−0.41*	−0.03	—													
11. Other GR anal. skill	2.11	0.69	117	0.05	0.03	0.14	−0.11	−0.00	−0.08	−0.25*	0.41*	−0.17	−0.45*	—												
12. Other GR persuasiven	1.85	0.84	117	−0.15	0.10	0.00	0.10	0.01	0.15	−0.23*	0.06	0.19	−0.68*	0.34*	—											
13. Self NEO sensitivity	166.22	14.05	149	0.12	−0.04	−0.12	0.03	−0.06	−0.04	0.34*	−0.25*	−0.10	0.35*	−0.23*	−0.18*	—										

14. Self NEO anal. skill	174.80	16.58	149	−0.08	0.00	0.03	0.07	0.24*	0.09	−0.08	0.01	0.09	−0.15	0.19*	−0.00	0.19	—													
15. Self NEO persuasiven	113.50	11.22	149	−0.17*	0.04	0.04	0.01	0.15	0.14	−0.32*	0.03	0.33*	−0.24*	−0.00	0.26*	−0.18*	0.46*	—												
16. Other NEO sensitivity	166.42	23.43	50	0.16	0.09	−0.08	0.19	0.15	0.03	0.32*	−0.06	−0.31*	0.05	−0.23	0.17	0.42*	−0.06	−0.19	—											
17. Other NEO anal. skill	182.14	18.81	50	0.00	0.10	−0.03	−0.22	−0.02	−0.36*	0.06	−0.13	0.09	0.15	−0.35*	0.17	0.28	0.43*	0.17	−0.00	—										
18. Other NEO persuasiven	110.70	13.21	50	−0.03	0.17	0.26	−0.08	0.11	0.02	0.05	−0.10	0.07	−0.11	−0.23	0.35*	−0.02	0.10	0.32*	−0.03	0.33*	—									
19. Other SRI sensitivity	61.98	8.84	121	0.29*	0.09	−0.00	0.18	0.16	0.13	0.27	−0.23	−0.05	0.44*	−0.11	−0.37*	0.34*	0.05	−0.18	0.43*	0.01	−0.16	—								
20. Other SRI anal. skill	42.09	4.45	121	0.15	0.08	−0.10	0.02	0.06	0.02	−0.02	−0.01	0.04	0.05	−0.03	−0.02	0.18*	0.17	0.11	0.14	0.08	−0.02	0.60	—							
21. Other SRI persuasiven	53.55	7.32	121	0.06	−0.01	−0.13	−0.08	−0.00	0.04	−0.19	−0.03	0.25*	−0.25*	0.04	0.24*	0.01	0.15	0.27*	−0.06	0.02	−0.06	0.14	0.54	—						
22. Self SRI sensitivity	66.01	7.96	145	0.10	0.04	−0.03	0.12	0.09	0.03	0.21*	−0.16	−0.06	0.10	−0.03	−0.08	0.29*	0.18*	0.13	0.26	−0.04	0.18	0.32	0.17	−0.08	—					
23. Self SRI anal. skill	44.14	4.38	145	0.08	0.09	−0.04	0.13	−0.01	0.06	0.24*	−0.15	−0.10	0.06	−0.10	0.02	0.11	0.08	0.09	0.16	−0.23	0.00	0.25	0.27*	0.06	0.69	—				
24. Self SRI persuasiven	55.56	6.16	145	0.01	0.13	0.04	0.10	0.14	0.13	0.15	−0.15	−0.01	−0.04	−0.03	0.06	−0.04	0.04	0.10	0.10	−0.16	0.27	0.09	0.10	0.04	0.40*	0.44*				

AC, assessment centre exercise; GR, general rating; NEO, NEO PI; SRI, situation response inventory.
*$p < 0.05$.