

A Triadic Approach to the Construct Validity of the Assessment Center

The Effect of Categorizing Dimensions into a Feeling, Thinking, and Power Taxonomy

Nanja J. Kolk¹, Marise Ph. Born², and Henk van der Flier³

¹Berenschot, Utrecht, ²Erasmus University, Rotterdam, ³Vrije Universiteit, Amsterdam, The Netherlands

Keywords: Assessment centers, dimensions, construct validity, taxonomy

Abstract: This study examined the influence on construct validity of implementing the triad Feeling, Thinking, and Power as a taxonomy for behavioral dimensions in assessment center (AC) exercises. A sample of 1567 job applicants participated in an AC specifically developed according to this taxonomy. Each exercise tapped three dimensions, one dimension from each cluster of the taxonomy. Confirmatory Factor Analysis of the multitrait-multimethod matrix showed evidence for construct validity. Thus, the ratings matched the *a priori* triadic grouping to a good extent. Practical implications are discussed.

Since AT&T's initial corporate application in the 1950s, the assessment center (AC) has thrived as one of the most popular methods for evaluating the performance of individuals for selection and development (e.g., Spychalski, Quinoñes, Gaugler, & Pohley, 1997). Due to the labor intensive nature of the AC architecture, practitioners are faced with numerous developmental and implementation problems. One of the problems highlighted by practitioners is how to select and define dimensions in exercises (e.g., Lievens & Goemaere, 1999). The fact that practitioners perceive problems with dimension selection and definition is consistent with the finding in the scientific AC literature that different dimensions within exercises correlate higher than similar dimensions across exercises, and that hence construct-related validity of the AC dimensions is low (e.g., Fleenor, 1996; Neidig & Neidig, 1984). Thus, from both practical and research perspectives, there is a need for a systematic procedure that enables AC developers to select independent and easily measurable dimensions, and that enables AC users to effectively distinguish between these dimensions.

The present study aims to contribute to AC practice and research by proposing a functional taxonomy of three broad dimension clusters, from which three operational dimensions are each time selected for every exercise. These clusters are Feeling (e.g., sensitivity, empathy, client orientation), Thinking (e.g., problem analysis, creativity, judgment), and Power (e.g., persuasiveness, risk taking, tenacity). The concepts of feeling, thinking, and power should be regarded as category labels for clusters of behavioral dimensions. This study investigates whether applying this taxonomy in a working AC as a means of selecting dimensions for each exercise increases construct validity.

A Taxonomy for Dimension Selection

Most AC dimensions are not completely orthogonal, which makes it difficult for assessors to decide which behaviors go with which dimensions. AC dimensions have been hypothesized to have a(n unknown) magni-

tude of true inter-correlation which lowers discriminant validity (Sackett & Dreher, 1982). In a similar vein, in order to obtain construct validity, care should be taken to define dimensions in a nonambiguous and unidimensional manner (Joyce, Thayer, & Pond, 1994; Kleinmann, Exler, Kuptsch, & Köller, 1995). The use of a large number of dimensions to be rated in an exercise has been viewed as causing cognitive overload and thus lowering construct validity (Bycio, Alvares, & Hahn, 1987; Gaugler & Thornton, 1989; Klimoski & Brickner, 1987; Silverman, Dalessio, Woods, & Johnson, 1986). In order to compensate for this cognitive overload, it seems that assessors themselves reduce a large number of dimensions into a smaller, manageable number of categories during the rating process (Sackett & Hakel, 1979; Sagie & Magnezy, 1997; Shore, Thornton, & MacFarlane Shore, 1990).

For these reasons, it has been suggested that the number of dimensions be reduced to three to five per exercise (Arthur Jr., Anthony Day; Gaugler & Thornton, 1989; McNelly, & Edens, 2001). While some studies have scrutinized the effects of conceptual distinctiveness and the number of AC dimensions on construct validity (see Lievens & Conway, 2001, for an overview of these studies), it has not yet brought about a functional taxonomy for AC development, and more specifically, for dimension selection.

Previous work in the fields of ACs, leadership, and personality jointly provides a small set of higher order constructs that meet the requirements of conceptual distinctiveness and nonambiguity and that could be used as an easily applicable taxonomy for selecting dimensions in exercises. Taking a step back in history, Plato, in *The Republic*, distinguished three classes of individual differences: Cognitive, the faculty of knowing, affective, the faculty of feeling, and conative, the faculty of volition (in Ackerman & Humphreys, 1993). Ackerman and Humphreys (1993) translated these faculties into modern terminology as *mental abilities*, such as intelligence or spatial ability, *temperament*, such as emotionality, objectivity, or masculinity-femininity, and *motivation*, such as effort. More contemporary studies in the relevant areas of ACs, leadership, and personality, pointed largely in the same direction. In the leadership literature, Yukl (1998) noted that managers use a threefold combination of leadership behavior: soft, rational, and power tactics. Also, Zand (1997) suggested that the three forces of "trust", "knowledge," and "power" coherently serve as the basis of effective leadership performance. Within the personality literature, we are most interested in the factors that describe the domain of interpersonal behavior because this may be of particular importance for the AC. Within this domain, Wiggins (1973, p. 479) noted that: "The circular arrangement of variables . . . was postulated to be

generated by two orthogonal axes representing Power (dominance versus submission) and Affiliation (love versus hate)."

These publications led to the notion that managerial behavior that is important in AC exercises may be effectively summarized in the three categories: Feeling (trust, affiliation, soft tactics, etc.), Thinking (knowledge, rational tactics, etc.), and Power (also previously labeled as such).

Within the AC construct validity literature, some researchers have focused on dimension categorization. It has been acknowledged that applicants are evaluated on the basis of their cognitive and interpersonal behaviors (Sagie & Magnezy, 1997; Shore et al., 1990). Shore et al. (1990) made this dual category distinction within 11 operational dimensions, between interpersonal style and performance style dimensions. This resulted in improved construct validity of the two dimension categories, compared to the construct validity of the operational dimensions. The present study follows up on Shore et al. (1990) by directly evaluating whether assessors are able to distinguish between more than two categories. The rationale for this is that assessors might be able to discern two elements within Shore's performance style dimensions, namely Thinking (e.g., planning & organizing, judgment, analytical skills, or Shore's "recognizing priorities") and Power (e.g., persuasiveness, assertiveness, decisiveness, or Shore's "work drive"). This corresponds to the AC dimension distinction proposed by Jansen (1991) between "thinking power" and "social power". The above mentioned dyadic approaches (Wiggins, Shore, Jansen) seem to complement each other: Concepts that are central in one approach are ignored in another approach and vice versa. The triadic approach can therefore be regarded as an integration of previous dyadic approaches.

Concluding, it seems reasonable to expect that behavioral dimensions within AC exercises can be categorized into a threefold taxonomy: feeling, thinking, and power, and that this leads to enhanced construct validity. This paper reports the data that result from implementing this triad in an operational AC. In this AC, each exercise consists of three operational dimensions, one dimension from the cluster Feeling, one from Thinking, and one from Power.

Method

Participants

Participants were 1567 Dutch job applicants (1079 male), tested in 1999 at a psychological consulting firm.

Table 1. Division of currently used AC dimensions into triad feeling–thinking–power.

Exercises	Feeling	Thinking	Power
Case analysis	Sociability	Analytical skills	Tenacity
Coaching interview	Sensitivity	Judgment	To give direction
Client interview	Client orientation	Judgment	Tenacity
Subordinate interview	Sensitivity	Judgment	Tenacity
Staff meeting	Cooperation	Judgment	Tenacity
In-basket exercise	Sensitivity	Analytical Skills	Delegating and control

The participants applied for a variety of jobs, mostly in management. Their mean age was 36 with a standard deviation of 8.

Assessors

The applicants were assessed by 26 assessors and 22 role-players. Both groups received an extensive and recurring assessor training and rated applicants on a day-to-day basis. The rater-ratee ratio was 2:1. Assessor and role-players were confronted with each applicant only once in order to make sure that the exercises were independent. This procedure minimizes rater bias between exercises (Kolk, Born, & van der Flier, 2002). After completion of each exercise, the role player and the assessor independently rated the applicants. Inter-rater reliability of the ratings in the present study was obtained by calculating the mean PPM correlation coefficient across dimensions and exercises, reaching a value of $r = .63$. Although this value indicates a moderate inter-rater reliability, it does not deviate from previously reported reliabilities (Thornton, 1992; Thornton & Byham, 1982).

Exercises

Each applicant participated in two exercises. These exercises were not the same for each applicant, as they applied for different jobs. Several types of commonly administered exercises were used (Thornton, 1992): interview simulations with subordinates, clients, and colleagues (68%), case-analyses (29%), and in-basket exercises (3%)*.

Dimensions

The exercises were designed to tap three dimensions, one dimension from each of the three clusters (Feeling,

Thinking, and Power). In other words, the dimensions were *a priori* grouped within the three clusters. This *a priori* categorization was based on two prestudies. First, we asked a group of 25 psychologists to sort approximately 350 behavioral examples into the categories Feeling, Thinking, Power or none of those. In order to ensure maximum conceptual dissimilarity, we used only those behavioral examples that fell into just one of the three categories (inter-rater reliability $> .80$) to create descriptions of the dimensions. Subsequently, we asked another team of four expert raters (psychologists with a mean rating experience of 14.5 years) to independently classify these dimensions, including their matching behavioral descriptions, into Feeling, Thinking, Power or none of those. This categorization procedure was done on rational grounds, following the Shore et al. (1990) and the Sagie and Magnezie (1997) studies.

Assessors rated the dimension in the AC on a 1 (*low*) to 5 (*high*) point scale, where ratings on intermediate scores (1.5, 4.8, etc.) were allowed. Table 1 shows the dimensions that we used per exercise, clustered into the proposed triad.

Analyses

Thus, the analyses were conducted on a conglomerate sample of exercises and dimensions to see if the intended dimensional triadic structure would emerge. For a formal test of the hypothesis, confirmatory factor analysis (CFA) was performed. The multitrait-multimethod (MTMM) covariance matrix was analyzed with LISREL 8.30 (Jöreskog & Sörbom, 1989). A commonly noted problem in analyzing trait by method MTMM matrices using CFA, is the occurrence of ill-defined solutions, such as convergence problems, negative (error) variances, or out-of-range factor intercorrelations (Kenny & Kashy, 1992). As in the majority of studies examining MTMM data (e.g., Lance et al., 2000), we ran into esti-

* By adding up several types of exercises, there could be more common variance shared by the dimensions than by the exercises, due to for instance a difference in exercise “difficulty”. This might result in stronger dimension factors. A multivariate analysis of variance on the dimension scores, using type of exercise as the independent variable, revealed a significant yet small effect (Pillai’s Trace for Exercise 1 and 2, respectively .09/.10; F: 10.81/12.05, $p < .05$). So, only 3% of the variance was accountable for a difference in exercise “difficulty.”

mation problems using the traditional CFA approach when testing some of the competing models. Kenny and Kashy (1992) suggested an alternative to testing the traditional CFA model that is not subject to the aforementioned problems: The so-called correlated uniqueness (CU) model. This approach specifies trait factors and does not create method factors, but allows its unique factors to be correlated across measures within the same method. Variances of the two methods (i.e., the exercises) were equalized throughout the models, for these can be assumed to be roughly similar.

Criteria for evaluating the competing CU models are firstly, measures of overall fit, namely the χ^2 /degrees of freedom ratio (should approach 1), the χ^2 *p*-value (should not be significant), the Adjusted Goodness-of-Fit Index, which adjusts the degrees of freedom relative to the number of variables in the model (AGFI, should approach 1), and the Root Mean Square Error of Approximation which evaluates the closeness of fit given the number of degrees of freedom (RMSEA, should be lower than .05). Secondly, since we were interested in comparing several competing models, we used Akaike's Information Criterion, which penalizes for leniency (the model with the smallest AIC should be selected).

The competing CU models were interpreted following the Widaman (1985) procedure of comparing the fit of hierarchically nested MTMM models (more specifically: Model E, see also Marsh, 1989). This means that a more parsimonious and, therefore, more restrictive model is tested against a less restrictive model using a likelihood ratio test. Generally, the more parameters there are to be estimated in a model, the better the model fits. Therefore, a less restrictive model can only be accepted when it provides a statistically significant improvement in the description of the data. CFA models based on MTMM data using this procedure always have a fixed method/trait factor intercorrelation null matrix, which is to say that trait and method factors are orthogonal (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Kenny & Kashy, 1992; Widaman, 1985). In the present study, the most restrictive yet meaningful model is a method-factor only model (i.e., correlated uniquenesses), for previous research has consistently shown the appearance of these method-factors (without trait-factors). Less restrictive models add parameters, until all meaningful parameters are estimated in the complete trait by method CU model.

A method-factor only model represents a primarily halo or exercise effect, indicating that assessors do not distinguish between any of the dimensions (Model I). This is the typical AC model that has usually been found in previous research. Secondly, a model is tested which adds only one general dimension factor (Model II). A model with two method- (correlated uniquenesses) and two trait-factors (Feeling and Power/Thinking) repre-

sents the hypothesized structure by Shore et al. (1990) (Model III). The complete trait by method CU model represents the intended triadic structure (Model IV). A χ^2 likelihood ratio test was used to determine whether the complete model (Model IV) fits the data significantly better than models nested within the complete model (Model I, II, and III).

Results

The MTMM correlation matrix of the two exercises is reported in Table 2. The observed correlation pattern does not meet the Campbell and Fiske (1959) MTMM criterion for establishing construct validity, for the heterotrait-monomethod correlations exceed monotrait-heteromethod correlations, instead of vice versa. This result was found in all previous AC construct validity studies, except one (i.e., Reilly, Henry, & Smither, 1990).

Table 3 shows the fit indices of the CFA of the four competing CU models. The method-factor only model (Model I: halo or exercise effect) did not show an adequate fit in the present study. In addition, Model II adds a general dimension factor. The χ^2 value of this model is also highly significant and the model does not fit well. Next, we tested a two-traits by two-methods model conforming to the Shore et al. (1990) assumption, where Feeling falls into the interpersonal style category, and Power and Thinking in the performance style category (Model III). This model also yields a significant χ^2 , and the fit indices are unsatisfactory. Differences in χ^2 /d.f., RMSEA, AGFI, and AIC magnitude all indicate a superiority of the complete three-traits by two-methods CU model (Model IV).

A χ^2 likelihood ratio test between the complete trait by method Model IV and most restrictive Model I (halo), indicates a significant improvement in fit ($\Delta \chi^2 = 192.54$,

Table 2. Multitrait-multimethod matrix.

	1	2	3	4	5	6
Exercise 1						
1. Feeling	1.00					
2. Thinking	.55	1.00				
3. Power	.38	.62	1.00			
Exercise 2						
4. Feeling	.22	.17	.16	1.00		
5. Thinking	.18	.21	.20	.54	1.00	
6. Power	.17	.16	.27	.40	.62	1.00

Note. All correlations are significant at the .01 level. Cases are excluded listwise. *N* = 1567. Mean heterotrait-monomethod correlation: *r* = .52; mean monotrait-heteromethod correlation: *r* = .23.

Table 3. Fit indices for the correlated uniqueness models.

Models		χ^2	d.f.	$\chi^2/d.f.$	RMSEA	AGFI	AIC
I.	2 orthogonal method correlated errors	195.02*	9	21.67	.11	.91	219.02
II.	2 orthogonal method correlated errors + 1 general dimension factor	50.67*	6	8.44	.07	.96	80.67
III.	2 orthogonal method correlated errors + 2 oblique trait-factors (Shore et al.)	36.47*	5	7.94	.06	.97	68.47
IV.	2 orthogonal method correlated errors + 3 oblique trait-factors	2.48	3	.83	.00	1.00	38.48

Note. $N = 1567$; * $p < .001$

6; $p < .001$). Similarly, the complete model also provided a significantly better description of the data than Model II ($\Delta \chi^2 = 48.19, 3; p < .001$), and Model III, representing the Shore et al. (1990) distinction ($\Delta \chi^2 = 33.99, 2; p < .001$). In sum, the hypothesized complete trait by method model provided the best description of our data.

As stated before, we analyzed a data set consisting of a conglomerate sample of dimensions and exercises, in order to investigate whether the intended triadic structure was indeed established. In addition, we performed CFA on a subsample of candidates participating in two *similar* exercises, tapping the same dimensions (i.e., two interview simulations, $n = 560$). Results indicated again a significantly better fit for the complete two by three factor model ($\chi^2 = 0.23, 3; p = .97$; RMSEA = .00), compared to Model I ($\Delta \chi^2 = 95.74, 6; p < .001$), Model II ($\Delta \chi^2 = 38.17, 3; p < .001$), and Model III ($\Delta \chi^2 = 22.77, 2; p < .001$).

Table 4 presents the completely standardized solution of the actual factor loadings, factor variances and intercorrelations, and the error uniquenesses of the CU model (Model IV). Within this model, evidence for discriminant validity is established by the interfactor correlations, which can be found in the lower half of Table 4. All interfactor correlations were significant, implying that discriminant validity coefficients are still poor. Evidence for convergent validity is established if the values in the standardized matrix of the factor loadings are significant (upper left corner of Table 4), which is the case.

Kenny and Kashy (1992, p. 170) noted that the correlated uniqueness approach assumes zero method-method correlations. When this assumption is not met, it can have a biasing effect on construct validity, through artificially enhancing convergent validity and worsening discriminant validity. This biasing effect may also be present in our model, for a zero method-method correlation is quite untenable in the case of ACs. This is not to say that the results from Table 3 will be negatively affected, as the fit would probably increase by adding a method-method in-

Table 4. Parameter estimates for the complete trait by method CU model (Model IV).

Source	Factor Loadings			Uniquenesses		
	7	8	9	1	2	3
Exercise 1						
Dimension 1	.68*	0	0	.79*		
Dimension 2	0	.64*	0	.39*	.81*	
Dimension 3	0	0	.60*	.22*	.46*	.75*
Exercise 2						
Dimension 4	.68*	0	0	.77*		
Dimension 5	0	.65*	0	.35*	.77*	
Dimension 6	0	0	.60*	.22*	.42*	.70*
Factor Intercorrelations						
	7	8	9			
7. Feeling	1.00					
8. Thinking	.82*	1.00				
9. Power	.69*	.76*	1.00			

Note. Values of 1 and 0 were fixed *a priori*. * $p < .05$

tercorrelation*. The results in Table 4, on the other hand, should be regarded with some caution, as both the factor loadings and the factor intercorrelations may be overestimated, at the expense of discriminant validity.

To sum up, on a matrix level, the present data show evidence for both discriminant and convergent validity – the complete model provides the best description of our data. On a parameter level, on the other hand, evidence for convergent validity is established, whereas evidence for discriminant validity seems weak.

Discussion

The results of the present study confirm previous findings showing that heterotrait-monomethod correlations (discriminant validity coefficients) are predominantly higher than monotrait-heteromethod correlations (convergent validity coefficients) (e.g., Sackett & Dreher, 1982; Nei-

* Because of these difficulties, we also performed a traditional correlated trait-correlated method CFA in addition to the correlated uniqueness model. (This analysis was possible within the complete sample ($N = 1567$), but not in the extracted subsample ($n = 560$), because the solution did not converge.) CFA of the correlated trait-correlated method revealed similar results as the correlated uniqueness model (Model IV: χ^2 : 3.06, 3, RMSEA: .00; AGFI: 1.00; AIC: 39.06). Further details of this analysis are available from the first author.

dig & Neidig, 1984; Bycio, et al., 1987; Klimoski & Brickner, 1987). Notwithstanding, the results extend previous research since they also support our *a priori* dimension grouping into the categories Feeling, Thinking, and Power. Confirmatory factor analysis yields a complete two-methods by three-traits model, providing convincing evidence for construct validity. The complete trait by method model fitted the data better than any alternative model. The method-factor-only model (halo or exercise effect), which is predominantly found in several previous studies (e.g., Chan, 1996; Fleenor, 1996; Joyce et al., 1994; Schneider & Schmitt, 1992), received no support in the present study. Sagie and Magnezy (1997) and Kudish, Ladd, and Dobbins (1997) found comparable CFA results. Additionally, a single dimension factor did not provide an adequate description of our data. In this light, our attempt to increase construct validity by applying the triadic taxonomy Feeling, Thinking, and Power to AC dimensions seems viable.

Moreover, our taxonomy seems to be a tenable extension of the Shore et al. (1990) distinction between interpersonal and performance style dimensions. Although this dual category proposition also received support in previous research (Sagie & Magnezy 1997), this study shows that assessors are able to distinguish between three orthogonal categories. Building an AC upon the dual taxonomy could perhaps even imply a loss of information, thereby impairing criterion-related validity (Lievens, 1998, p. 146). Therefore, in designing an AC, a triadic approach might be more legitimate.

Implications for AC Practitioners

This study provides an easily applicable tool for AC practice, through facilitating the selection of maximally dissimilar dimensions within exercises and by improving construct validity of these dimensions. In our own practical experience, the taxonomy proved a helpful tool in formulating conceptually dissimilar dimensions, and thereby facilitating the rating task. In addition, providing feedback in terms of Feeling, Thinking, and Power resulted in greater comprehension and acceptance by assessors, as well as applicants. Receiving feedback on these three domains makes it easier to direct future development, than receiving such feedback on all dimensions individually.

The taxonomy presented in this study may be used by AC developers to select and define dimensions for exercises. Preexisting dimensions or behavioral indicators that follow from job analysis may be categorized in the triad by (expert) raters during AC development workshops. After determining which dimensions are to be measured, each dimension can be attributed to one of the

three clusters, feeling, thinking, and power. Subsequently, exercises can be developed that tap these dimensions. Ahmed, Payne, and Whiddett (1997) provided a model for exercise design that may be of help.

If a target dimension cannot be categorized in one of the three clusters, it may indicate that the dimension is in fact a multi-faceted concept (e.g., the dimension "leadership"). If so, the dimension may also be difficult to evaluate during an exercise, because the assessors may not agree on the meaning of the dimension, or they may not know which behaviors to look for. A solution might be to divide this dimension into two or three parts, that can be attributed to Feeling, Thinking or Power. For instance, "leadership" may be divided into the dimensions delegation (thinking), decisiveness (power), and sensitivity (feeling), depending on the definition of leadership that derives from the job analysis.

Another solution to overcome the commonly noted confusion over the meaning of the target dimensions is a so-called frame-of-reference training, which has been recently advocated (Lievens, 2001; Schleicher & Day, 1998). This type of training aims to provide a shared performance theory for raters, such that each rater evaluates applicants on the basis of the same conceptualization of effective performance. This is generally established during preassessment workshops. The Feeling, Thinking, Power triad may provide a helpful contribution to this type of workshop, in that assessors are not only trained on a shared frame-of-reference, but also on means to distinguish the dimensions during AC exercises. Practical implications may also involve adjustments in AC training programs, in which assessors are trained to classify observed behaviors as Feeling, Thinking, and Power.

One possible pitfall of using a smaller number of dimensions is defining them at too high level of abstraction in order to cover a broad scope of behavior (the information "loss" caused by the reduction of the number of dimensions is compensated for by giving the remaining dimensions a broad definition). Lievens and Conway (2001) warned that when broadly defined dimensions are used behaviors may overlap between those dimensions making it difficult for assessors to distinguish among them. The feeling, thinking, power taxonomy should not be misused to define broader dimensions. The taxonomy has only one purpose, and that is to carefully *select* dimensions – that have already been specifically defined – from each of the clusters feeling, thinking, and power.

Limitations and Directions for Future Research

A first limitation of our study is that we used no more than two exercises to represent a full AC, while in prac-

tice an AC usually includes a broader range of different exercises (perhaps five).

Another potential shortcoming in the present study is that the dimensions measured in the exercises were often similar within the three domains (Table 2). For instance, the dimensions Judgment, Tenacity, and Sensitivity were used in all but three exercises. It is conceivable that using more diversified dimensions might influence the results. In addition, the clustering of dimensions into the triad was done on an *a priori* basis, using expert raters for classifying the dimensions. Results indicated that implementing a triadic taxonomy indeed yields a meaningful triadic latent within and across exercise structure. Thus the ratings match the *a priori* triadic grouping to a good extent. Yet, we did not discover whether the Feeling, Thinking, and Power domains can be adequately and fully measured within an AC. As such, the current results should be regarded as an incentive for testing this taxonomy. Future research could make an attempt to reanalyze previous research through meta-analysis, and cluster multiple dimensions within the presently proposed taxonomy. Subsequently, structural equation models could be tested in order to confirm the clustering into the triadic taxonomy, where one-factor models could be fitted on the clustered dimensions. However, the often lacking description of the complete MTMM matrix on a dimension level may present a difficulty for such a reanalysis.

Another route to test whether the triadic approach is at all feasible and perhaps even superior to a dyadic approach, is to experimentally vary the dimension composition per exercise, while holding all boundary conditions constant (e.g., in one condition the dimensions are selected according to the feeling/thinking/power taxonomy, while in another condition three dimensions are randomly selected)*.

Acknowledgments

We gratefully acknowledge our colleagues at LTP Amsterdam for making this research possible, sharing their data and ideas with us, as well as providing us with financial means. We also thank Filip Lievens for his helpful comments on an earlier version of this paper.

References

- Ackerman, P.L., & Humphreys, L.G. (1993). Individual differences in industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 223–282). Palo Alto, CA: Consulting Psychologists Press.
- Ahmed, Y., Payne, T., & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment*, 5, 62–68.
- Arthur Jr., W., Anthony Day, E., McNelly, T.L., & Edens, P.S. (2001, April). *The criterion-related validity of assessment center dimensions: Distinguishing between methods and constructs*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology Conference, San Diego, CA.
- Bycio, P., Alvares, K.M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology*, 72, 463–474.
- Campbell, J.P., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Chan, D. (1996). Criterion and construct validation of an assessment center. *Journal of Occupational & Organizational Psychology*, 69, 167–181.
- Donahue, L.M., Truxillo, D.M., Cornwell, J.M., & Gerrity, M.J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality*, 12, 85–108.
- Fleener, J.W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business & Psychology*, 10, 319–335.
- Gaugler, B.B., & Thornton, G.C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, 74, 611–618.
- Jansen, P.G.W. (1991). *Het beoordelen van managers. Effectiviteit van assessment center methoden bij selectie en ontwikkeling van managers* [Management assessment: Effectiveness of assessment center methods for selecting and developing managers]. Baarn: Uitgeverij H. Nelissen.
- Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS Inc.
- Joyce, L.W., Thayer, P.W., & Pond, S.B. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology*, 47, 109–121.
- Kenny, D.A., & Kashy, D.A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165–172.
- Kleinmann, M., Exler, C., Kuptsch, C., & Köller, O. (1995). Unabhängigkeit und Beobachtbarkeit von Anforderungsdimensionen im Assessment Center als Moderatoren der Konstruktvalidität [Independence and observability of assessment center dimensions as moderators of construct validity]. *Zeitschrift für Arbeits- und Organisationspsychologie*, 39, 22–28.
- Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, 40, 243–260.
- Kolk, N.J., Born, M.Ph., & Van der Flier, H. (2002). Common rater variance as explanation for the lack of construct validity of assessment center dimensions. *Human Performance*.
- Kudisch, J.D., Ladd, R.T., & Dobbins, G.H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior & Personality*, 12, 129–144.
- 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.

* We thank an anonymous reviewer for this suggestion.

- Lievens, F. (2001). Assessor training strategies and their effects on accuracy, inter-rater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255–264.
- 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.
- Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141–152.
- Lievens, F., & Goemaere, H. (1999). A different look at assessment centers: Views of assessment center users. *International Journal of Selection and Assessment, 7*, 215–219.
- Neidig, R.D., & Neidig, P.J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182–186.
- Reilly, R.R., Henry, S., & Smither, J.W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71–84.
- Sackett, P.R., & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401–410.
- Sackett, P.R., & Hakel, M.D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. *Organizational Behavior & Human Performance, 23*, 120–137.
- Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment center construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103–108.
- Schleicher, D.J., & Day, D.V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76–101.
- 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.
- Shore, T.H., Thornton, G.C., & MacFarlane Shore, L.M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101–116.
- Silverman, W.H., Dalessio, A., Woods, S.B., & Johnson, R.L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39*, 565–578.
- Spychalski, A.C., Quinoñes, M.A., Gaugler, B.B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50*, 71–90.
- Thornton, G.C. (1992). *Assessment centers in human resource management*. Reading, Massachusetts: Addison-Wesley.
- Thornton, G.C.I., & Byham, W.C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.
- Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.
- Wiggins, J. (1973). *Personality and prediction: Principles of personality assessment*. Reading, Massachusetts: Addison-Wesley.
- Yukl, G. (1998). *Leadership in organizations* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Zand, D.E. (1997). *The leadership triad: Knowledge, trust, and power*. Oxford: Oxford University Press.

Nanja J. Kolk
 Berenschot Business Management – The Change Factory
 Europalaan 40
 NL-3526 KS Utrecht
 The Netherlands
 Tel. +31 30 2916916
 Fax +31 30 2947090
 E-mail N.Kolk@Berenschot.com
