Journal of Applied Psychology
2006, Vol. 91, No. 3, 555–566


0021-9010/06/$12.00 DOI: 10.1037/0021-9010.91.3.555

# A Construct-Driven Investigation of Gender Differences in a Leadership-Role Assessment Center

Neil Anderson
University of Amsterdam

Filip Lievens
Ghent University

Karen van Dam
Tilburg University

Marise Born
Erasmus University

This study examined gender differences in a large-scale assessment center for officer entry in the British Army. Subgroup differences were investigated for a sample of 1,857 candidates: 1,594 men and 263 women. A construct-driven approach was chosen (a) by examining gender differences at the construct level, (b) by formulating a priori hypotheses about which constructs would be susceptible to gender effects, and (c) by using both effect size statistics and latent mean analyses to investigate gender differences in assessment center ratings. Results showed that female candidates were rated notably higher on constructs reflecting an interpersonally oriented leadership style (i.e., oral communication and interaction) and on drive and determination. These results are discussed in light of role congruity theory and of the advantages of using latent mean analyses.

*Keywords:* assessment centers, gender differences

The examination of gender differences in assessment centers has a long research tradition (Baron & Janman, 1996). The general conclusion from prior studies is a mixed one. Empirical studies are approximately evenly split between studies showing no significant differences between men and women and studies indicating that women scored somewhat higher than men. Unfortunately, the majority of past studies have examined gender differences only at the level of the overall assessment center rating (OAR), leaving open to doubt whether differences on particular dimensions, or constructs, led to final OAR differences. The findings of all previously published studies on gender differences in assessment centers are summarized in Table 1.

The aim of this study was to conduct a construct-driven examination of gender differences in assessment centers. This means that the present study focused on gender differences at the level of the assessment center constructs instead of at the more diffuse level of the OAR. We also aimed to advance prior research on gender differences in assessment centers in two other ways. First, from a substantive point of view, we generated a priori hypotheses about gender differences on these constructs on the basis of theoretical and empirical research on gender differences in leadership style (e.g., Eagly, 1987; Eagly & Johnson, 1990; Eagly & Karau, 2002). Second, from a methodological point of view, we used both effect size statistics and latent mean analyses to investigate gender differences in assessment center ratings. Specifically, male–female subgroup differences were examined in terms of both the observed (thus error-laden) dimension ratings by assessors and the latent constructs. The following section discusses each of these contributions in more detail. The setting of this study is a large-scale assessment center for officer entry in the British Army.

## Study Background

### *The Construct-Driven Perspective in Assessment Centers*

A first critical limitation of most of the studies in Table 1 is that gender differences were typically evaluated at the level of the OAR. Although it is clear that the OAR is of great practical importance (hiring decisions are contingent on it), it typically is a summary rating of assessor evaluations on a wide variety of dimensions (e.g., from more cognitive-oriented dimensions to more interpersonally oriented dimensions) in a very diverse set of simulation exercises (e.g., individual exercises, one-on-one exercises, group exercises; Howard, 1997; Schneider & Schmitt, 1992; Tett & Guterman, 2000; Zedeck, 1986). The fact that the OAR is such an amalgam of various ratings reduces its conceptual value (Arthur, Day, McNelly, & Edens, 2003) and might explain the inconsistent results found in prior gender research in assessment

Neil Anderson, Department of Work and Organizational Psychology, University of Amsterdam, Amsterdam, The Netherlands; Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Brussels, Belgium; Karen van Dam, Department of Psychology, Tilburg University, Tilburg, The Netherlands; Marise Born, Institute of Psychology, Erasmus University, Rotterdam, The Netherlands.

Neil Anderson and Filip Lievens contributed equally to this article, and so order of authorship is alphabetical.

An earlier version of this article was presented at the European Congress of Work and Organizational Psychology, Lisbon, Portugal, May 2003. We thank Frederik Anseel, Agneta Fischer, Jesus Salgado, and Annelies van Vianen for their comments on drafts of this article.

Correspondence concerning this article should be addressed to Neil Anderson, who is now at the University of Amsterdam Business School, Roetersstraat 11, 1018WB Amsterdam, The Netherlands. E-mail: n.r.anderson@uva.nl

Table 1
*Overview of Previous Published Studies of Gender Differences in Assessment Centers*

| Study | Sample | Sample characteristics | Findings |
|---|---|---|---|
| Moses (1973)* | 85 (39 m, 46 f) | Early identification assessment center | No significant difference in promotion ratios |
| Moses & Boehm (1975)* | Unspecified | Applicants for managerial positions | Distribution of ratings was similar for men and women |
| Alexander, Buck, & McCarthy (1975) | 111 (not available) | Applicants for jobs in Federal Administration | No significant difference in promotion ratios |
| Schmitt & Hill (1977) | 306 (184 m, 122 f) | Applicants for supervisory positions | Small differences between men and women |
| Ritchie & Moses (1983) | Unspecified | Applicants for supervisory positions | Similar percentages of men and women possessed middle-management potential |
| L. R. Anderson & Thacker (1985) | 64 (49 m, 15 f) | Applicants for computer sales positions | No significant difference in OAR |
| Walsh, Weinberg, & Fairfield (1987) | 1,035 (817 m, 218 f) | Salespersons in financial services | Significant difference in OAR in favor of women |
| Shore (1992) | 436 (375 m, 61 f) | Early career managers | Significant difference in favor of women only on performance-style skills |
| Schmitt (1993) | 2,910 (1,481 m, 1,350 f) | Applicants for school administrators | Significant performance differences favoring women on all dimensions |
| Weijerman & Born (1995) | 77 (38 m, 39 f) | Policy advisors in the Netherlands | Significant difference in favor of women only in a role play |
| Bobrow & Leonards (1997) | 169 (not available) | Supervisors in customer services | Women scored 0.20 *SD* higher on OAR |
| Shore, Tashchian, & Adams (1997) | 209 (119 m, 90 f) | Employees of financial services organization | No significant difference between men and women on OAR |

*Note.* The studies with an asterisk investigated the criterion-related validity of assessment centers. However, both studies report that checks for male–female differences in assessor ratings were carried out on unspecified additional data sets. Although no details of these data sets are given, both studies report that similar distributions were observed. m = male; f = female; OAR = overall assessment center rating.

centers. Furthermore, evaluation of subgroup differences at the level of the OAR may conceal significant differences that exist at the level of dimensions.

In fact, recent assessment center research has explicitly recognized that assessment centers should not be treated as a monolithic entity (Arthur et al., 2003; Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Lievens & Conway, 2001). Lievens and Conway (2001) reviewed 34 studies on assessment center construct validity and concluded that dimensions as constructs deserve their place in assessment centers (see also Woehr & Arthur, 2003). In a study, Arthur et al. (2003) cogently argued that assessment centers are best conceptualized as a method that can be designed to measure a variety of constructs (dimensions). As a result, they posited that it does not make a lot of sense to evaluate the criterion-related validity of assessment centers at the level of the OAR. Instead, they assessed the criterion-related validity of assessment centers at the construct level. Their meta-analytical results noticeably showed that some dimensions (e.g., dimensions related to conscientiousness) demonstrated higher criterion-related validities than other dimensions. In a similar vein, Goldstein and colleagues (Goldstein et al., 1998; Goldstein, Yusko, & Nicolopoulos, 2001) used a construct-driven approach to explain the equivocal findings in the assessment center literature about race differences (see Hoffman & Thornton, 1997, for a review). Specifically, Goldstein and colleagues demonstrated that subgroup (Black–White) mean differences were a function of the cognitive loading of the simulation exercises. That is, when more cognitive-oriented exercises (e.g., in-basket exercise) or more cognitive-oriented dimensions (e.g., problem analysis, judgment) were used, there were stronger race differences.

Similar to Goldstein et al. (1998), recent research into other selection procedures has also emphasized the importance of distinguishing the method of measurement (e.g., tests, interviews, work sample) from the constructs being measured (e.g., cognitive ability, conscientiousness; Schmitt, Clause, & Pulakos, 1996). For example, Huffcutt, Conway, Roth, and Stone (2001) showed that subgroup differences in the interview varied according to the constructs measured. Other studies revealed that it is possible to reduce subgroup differences by using another method while holding the construct constant (Arthur, Edwards, & Barrett, 2002; Chan & Schmitt, 1997; Schmitt & Mills, 2001). Given these recent developments in linking subgroup differences to the constructs measured, this study similarly applied a construct-driven approach to scrutinize gender differences in assessment centers.

### *Gender Differences in Leadership Style*

A second drawback of prior research on gender differences in assessment centers is that prior studies explored whether gender differences occurred and provided post hoc explanations for the results obtained. A better strategy, however, consists of drawing on the voluminous literature on gender differences to specify a priori hypotheses. For instance, in this study, meta-analyses of gender differences on leadership dimensions served as fruitful inspiration for a better understanding of how male and female candidates are evaluated in assessment centers designed to identify leaders or future managers. The literature on differences on leadership dimensions is useful because the simulated setting of an assessment center designed to identify leaders parallels the design of many laboratory studies on gender differences in leadership.

Inspection of this meta-analytic research base about gender differences in leadership (Eagly & Johnson, 1990; Eagly & Karau, 1991; Eagly, Karau, & Makhijani, 1995) demonstrates that, in general, men and women are equally effective as leaders (see also Cleveland, Stockdale, & Murphy, 2000; Powell, 1993), although men and women seem to lead in different ways. That is, there is no significant performance level difference, but there are notable differences on leadership dimensions. This effect was noticed in a meta-analysis examining gender differences in the context of the emergence of leaders in groups that were initially without leaders (Eagly & Karau, 1991). Given its resemblance to assessment center situations, this specific meta-analysis can be used to inform our hypotheses with regard to male–female subgroup differences on the specific dimensions measured in this assessment center (see the Appendix).

Generally, Eagly and Karau's (1991) meta-analysis revealed stereotypical gender differences in task-oriented and interpersonally oriented leadership style. Men emerged more often as task-oriented leaders who displayed directive and controlling leadership styles. Other meta-analytic research found that male managers were more motivated to work in a competitive environment, exert an assertive role, impose their wishes on others, and stand out from the group (see also Eagly, Johannesen-Schmidt, & Van Engen, 2003; Eagly, Karau, Miner, & Johnson, 1994). These leadership differences also translated to leadership effectiveness, as men were more effective than women in roles that were defined in more masculine terms (Eagly et al., 1995). Moreover, these meta-analytic findings are generally supportive of role congruity theory (Eagly, 1987; Eagly & Karau, 2002). This theory posits that people are expected to engage in activities that are consistent with their gender roles. Violations of these gender stereotypes may lead to lower performance evaluations of women. Our hypotheses are therefore based on both of these meta-analytic findings and on role congruity theory. As shown in the Appendix,[1] both problem solving and impact can be considered important dimensions of a task-related leadership style. In fact, problem solving was defined in this assessment center as the ability to deal with problems in a no-nonsense manner, whereas impact was defined as the extent to which a person assumes the lead of a group and maintains control. Therefore, we proposed the following hypothesis:

> *Hypothesis 1*: There will be significant gender differences favoring male candidates on task-oriented leadership dimensions such as impact and problem solving.

Eagly and Karau's (1991) meta-analysis also revealed that women emerged more often than men as social leaders who facilitated interpersonal relations and contributed to good morale. A follow-up meta-analysis examined leader effectiveness, indicating that women were more effective than men in roles that were defined in less masculine terms (Eagly et al., 1995). Again, role congruity theory posits that women lead as social leaders as opposed to task leaders (Eagly, 1987; Eagly & Karau, 2002). As shown in the Appendix, interaction (the ability to relate to others individually and within groups) and communication (the ability to relate fluently to others) are dimensions measured in this officer assessment center. Both can be considered key components of an interpersonal leadership style. This led to the following hypothesis:

> *Hypothesis 2*: There will be significant gender differences favoring female candidates on interpersonally oriented leadership dimensions such as communication and interaction.

Counter to role congruity theory, exactly the opposite predictions can be derived from expectancy violation theory (Jussim, Coleman, & Lerch, 1987). This theory proposes that behaviors that violate gender stereotypes will be more positively evaluated. In other words, women will receive higher ratings on dimensions that do not conform to gender stereotypes (e.g., task-oriented leadership dimensions). Thus, expectancy violation theory provides a competing theoretical framework for both hypotheses formulated above. A subsidiary aim of this study was therefore to test whether the data from this operational assessment center supported gender differences in favor of role congruity theory or, alternatively, in favor of expectancy violation theory.

### *Testing for Gender Differences*

As previously noted, prior research on gender differences in assessment centers has primarily focused on mean score differences. These mean score differences were typically differences at the level of the OAR. The usual approach was to break down the OAR by gender and then compare ratings of male or female candidates using a *t* test or a similar procedure. If an observed score difference was found, the interpretation was that this was due to a gender difference. However, this is essentially a methodologically inadequate approach of testing for subgroup differences. By only evaluating subgroup differences on observed ratings (i.e., the dimensions actually rated by assessors), research runs the risk of misattributing differences due to a failure to consider the impact of measurement error (Cheung & Rensvold, 2000; Hattrup, Schmitt, & Landis, 1992). Along these lines, Hoyle and Smith (1994) concluded that a comparison between groups on the basis of observed mean differences alone on measures whose reliability and factorial validity have not been proven to be (at least partially) invariant across the groups can be misleading and is a classic example of "comparing apples and oranges" (p. 433; see also Ryan, Chan, Ployhart, & Slade, 1999).

Modeling latent means in structural equation modeling (also known under the aliases of *multiple-group mean and covariance structures analysis* or *structured means analysis*) provides researchers with a methodologically more refined approach for testing for differences across two groups (Cheung & Rensvold, 2000; Collins & Gleaves, 1998; Hancock, 1997; Hoyle & Smith, 1994; Little, 1997; Ployhart & Oswald, 2004). In particular, it enables researchers to assess measurement equivalence between groups prior to testing for latent group-mean differences. If measurement equivalence (or at least partial measurement invariance) can be established, this means that both the reliability (measurement error) and the factorial structure of the scores are held constant across groups. Hence, in case of measurement invariance, researchers should be able to determine whether there are gender differences on the (latent and error-free) factors of interest. In

[1] No hypotheses were formulated regarding the other dimensions (drive and determination and resistance to stress) because these dimensions could not be linked to the interpersonal versus task-oriented distinction made in meta-analyses about gender differences in leadership.

assessment centers, these factors pertain to both exercises and dimensions. So, one is able to test whether there are gender differences on the (latent and error-free) exercise and dimensions. As we conducted a construct-driven examination of gender differences in assessment centers, we focused on the latent mean differences on the dimensions. However, the fact that the structural equation modeling analyses tease out dimension variance from exercise variance is interesting because it helped us to evaluate whether there exist gender differences at the level of the dimensions (constructs).

## Method

### Sample

A sample of 1,857 candidates attending the British Army officer assessment center participated in the study. These candidates applied for the 2-year officer training program in the British Army. Of this total, 1,594 (85.8%) were male and 263 (14.2%) were female, with these percentages reflecting the overall numbers of male and female applicants for officer training in the British Army. Candidate ages ranged from 18 to 31, with a mean of 23.2 ($SD = 2.2$). The majority of candidates (over 80%) were either university graduates or were currently studying at a British university.

### Recruitment Procedure and Prescreening

Applicants were recruited into the selection process either by face-to-face contact with tri-service (army, navy, and air force) recruiting officers who were based in city center offices located throughout the United Kingdom or by the applicant logging onto the army's recruitment Web site and subsequently being directed to his or her local tri-service office. The army also sponsors a number of individuals through their university degree programs, with the agreement that individuals apply for entry into officer training upon graduation but with no guarantee of selection (such candidates typically also serve as part-time reservists during their degree studies).

The target applicant population consisted of students from all British universities. The recruitment process adhered to a long-standing equal opportunities policy whereby male and female applicants were assured of equal treatment during selection and military service. This policy had been in place since the mid-1970s in an attempt to open up the armed forces to an increasing number of female officers and basic grade recruits (see also Dobson & Williams, 1989). All applicants were prescreened on the basis of school results at age 18, expected degree classification, membership and leadership posts in extracurricular sports clubs and societies, and experience of the army as reservists during their degree programs. Prior to the assessment center, candidates were screened on various psychometric tests (e.g., cognitive ability).

Results of *t* tests conducted showed that there were no significant differences in terms of age, cognitive ability, and educational attainment between the male and female candidates who participated in the assessment center. This attests to the fact that our male and female samples were well matched on these variables.

### Overview of Assessment Center

Each assessment center took place during 3.5 days and was attended by 8 candidates. Data were collected from approximately 250 assessment centers over a 5-year period. The purpose of the assessment center was "to select from the field of candidates of acceptable education and physical standards, those with the potential qualities of character, ability and leadership who should after training be able to lead a platoon or troop in battle" (Dobson & Williams, 1989, p. 313).

The origins of the British Army officer assessment center go back to the Regular Commissions Board and the War Office Selection Boards (Vernon & Parry, 1949). Accordingly, the British Army officer assessment center is one of the oldest still existing assessment centers. It has considerable similarity with American armed forces assessment centers for officer training where leadership potential is likewise assessed (Borman, 1982). Prior research attests to the good predictive validity of this British Army officer assessment center, with validity coefficients for training and job performance in the .30s (Dobson & Williams, 1989).

### Assessment Center Dimensions and Exercises

In line with common assessment center practices, the assessment center did not measure an omnibus leadership dimension. Instead, the assessment center targeted eight subdimensions based on regular job analyses of a leadership position (i.e., the officer position) going back over several decades. Definitions of each of these leadership dimensions are presented in the Appendix.

With the officer position and assessment center dimensions in mind, we developed exercises. Exercise content was generated on the basis of job analysis information and input from subject matter experts. Eight exercises attempted to tap the dimensions of interest. The first exercise was a leaderless group discussion in which the candidates discussed during 40 min various subjects of topical interest. Next, candidates performed in so-called opening tasks. These opening tasks required the cooperation of the group for their completion. In general, they could be completed by physical efforts on the part of the group. Typically, after a task was explained to the candidates, the group worked out a possible solution and subsequently tried to complete the task. Apart from the leaderless group discussion and opening tasks, there were two planning exercises. The first was an individual written planning exercise in which each candidate considered a problem alone for 90 min and wrote out an individual plan. The second was a group planning exercise in which all candidates spent 15 min deciding on a group plan. The next two exercises were practical command tasks (own command tasks and other command tasks). First, each candidate in turn was briefed about the problem and was given 2 min to consider a plan to solve the problem. Second, each candidate in turn was placed in command of the group and undertook the solution of an outdoor practical task. Another exercise was a 5-min lecturette, which aimed to place the candidate in an officer role as an instructor. A candidate could choose to focus his or her talk on any one of five subjects chosen from his or her interests and background. At the end of each lecturette, the group (other candidates) asked questions and challenged the candidate. Finally, there was also a physical exercise (i.e., individual obstacles). Candidates had to complete as many obstacles as possible within 3 min. Generally, four of these exercises (own command tasks, other command tasks, opening tasks, and individual obstacles) were framed in a military context, whereas the remaining four exercises (planning exercise, group planning exercise, leaderless group discussion, and lecturette) were not framed in a military context.

In this study, we included in the analyses only exercises that measured more than one dimension and dimensions that were measured in multiple exercises (see Lievens & Conway, 2001). If we applied these inclusion criteria, two exercises (the individual obstacles exercise and the written planning exercise, respectively) were excluded, as they measured only one dimension. After excluding these two exercises, the dimension of analysis and planning was measured in only one exercise, and the dimension of physical ability was no longer measured. So, these two dimensions were also excluded from the analyses. So, in this study, six dimensions and six exercises were used in the analyses. The dimension–exercise matrix is presented in Table 2.

### Assessors, Training, and Rating Process

Experienced army officers served as assessors. All assessors had attended a training seminar that lasted for 3 days. Training content was quite comprehensive and included the explanation of dimensions and exercises used. In addition, the training focused on practice and feedback in the process of observing, recording, classifying, integrating, and reporting

Table 2
*Dimension by Exercise Matrix*

| Dimension | Leaderless group discussion | Opening tasks | Group planning exercise | Own command tasks | Other command tasks | Lecturette | Obstacles exercise | Written planning exercise |
|---|---|---|---|---|---|---|---|---|
| Oral communication | X | | X | | | X | | |
| Interaction | X | X | X | X | X | X | | |
| Problem solving | | X | | X | X | | | |
| Impact | X | X | X | X | | | | |
| Drive and determination | | X | | X | X | | | |
| Reaction to stress | | | X | X | | X | | |
| Analysis and planning | | X | | | | | | X |
| Physical ability | | | | | | | X | |

assessee behavior. At the end of the training track, assessors went themselves through the various exercises, then subsequently observed an assessment center throughout, before being authorized to act as assessors.

Assessors rated candidates on the dimensions using behaviorally anchored rating scales. These behaviorally anchored rating scales were derived on the basis of generally accepted practices (e.g., Champion, Green, & Sauser, 1988). In most exercises, two assessors rated the candidates. Intraclass correlations (ICC 2.1; Shrout & Fleiss, 1979) were computed and were satisfactory ($M = 5.65$). Consistent with current assessment center practice, participants were rated by different assessors across exercises.

After completion of all exercises, assessors met to discuss their observations and ratings with one another and agreed on an OAR. Ratings were combined using a complex multistage clinical integration process. This essentially involved reaching agreement initially over the combined rating for each candidate on each dimension by exercise, as per the targeted matrix given in Table 2. Assessors then discussed and agreed on a combined rating by dimension across all exercises measuring that particular dimension, most commonly equating to a rough mean rating by dimension (see Goffin, Rothstein, & Johnston, 1996). Assessors were able to challenge the agreed-upon rating at both stages, but most commonly agreement was reached by taking the midpoint between assessor ratings or, if this was not possible, by the senior officer on the panel proposing the final combined rating. A final board meeting then combined these ratings into an OAR with the following anchors: 1 = *not recommended*, 2 = *recommended as risk entry*, 3 = *recommended as minor risk entry*, and 4 = *recommended unconditionally*. All board assessors needed to agree on the final recommendation. Across the 1,857 candidates in this study, the distribution of OARs was as follows: 1 = 42.5% ($N = 785$), 2 = 23.6% ($N = 435$), 3 = 30.2% ($N = 558$), and 4 = 3.7% ($N = 68$).

## Statistical Analyses

*Test of fit of measurement models (within each group).* Prior to testing for measurement invariance (and subsequently latent mean differences), we started by testing several measurement models that represented different conceptualizations of assessment centers (see Lievens & Conway, 2001, for a review). Actually, there is considerable debate in the assessment center literature whether assessment centers actually measure the dimensions they are purported to measure. Therefore, it was important to examine the underlying structure of the ratings in terms of dimensions and exercises prior to testing for measurement invariance and examining our hypotheses about latent means (recall that our hypotheses were formulated at the level of the assessment center dimensions).

First, we tested a dimensions-only model. In this model, assessment centers were conceived as a way to measure stable traits (Sackett & Dreher, 1982). Hence, this model included a factor for each assessment center dimension but ignored exercises. Second, we tested an exercises-only model. This model can be thought of as the opposite of the dimensions-only model—it included a factor for each assessment center exercise but ignored dimensions. In this model, assessment centers were conceptualized to be nothing more than a series of miniaturized work samples of managerial behavior (Robertson, Gratton, & Sharpley, 1987). The third model represented the assumption that assessors were unable to distinguish among dimensions (see also Bycio, Alvares, & Hahn, 1987; Kudisch, Ladd, & Dobbins, 1997). This model consisted of exercise factors and one global dimension factor. The fourth model was a combination model containing both exercise and dimensions. This model reflected the inherent design of assessment centers, which are designed to measure dimensions in exercises (see also Bycio et al., 1987; Kudisch et al., 1997).

To test the fit of these measurement models through confirmatory factor analysis within each sample (male candidates and female candidates), we used EQS (Bentler, 1995) to derive maximum-likelihood estimates for the input covariance matrix. Automatic start values were used to fit each model, and if the model failed to converge after 250 iterations, the start values were set near estimates from other specifications that converged, and the analysis was repeated.

We used several fit indices to assess how these models represented the data. Absolute fit indices such as the $\chi^2$ statistic as well as incremental fit statistics such as the comparative fit index (CFI) and the root-mean-square error of approximation (RMSEA) were used. For the CFI, values greater than .95 constitute good fit, and values greater than .90 constitute acceptable fit (Medsker, Williams, & Holahan, 1994). For the RMSEA, it has been suggested that values less than .05 constitute good fit, values in the .05–.08 range constitute acceptable fit, values in the .08–.10 range constitute marginal fit, and values greater than .10 constitute poor fit (Browne & Cudeck, 1992).

*Tests of invariance of measurement model (stacked multiple groups).* Once an appropriate measurement model was established in each of the samples, we examined the invariance or equivalence of this measurement model across assessors' ratings of male versus female candidates as a prerequisite to comparing latent mean differences. To this end, we conducted multiple-group confirmatory factor analyses using EQS. As noted by Hoyle and Smith (1994), measurement invariance should be regarded as a continuum ranging from nonequivalence of the form of the measurement model (i.e., different number of factors account for ratings of male and female candidates) to equivalence of the form and all parameters of the measurement model (i.e., factor loadings, measurement errors, factor variances and covariances). Hence, to examine invariance of the measurement model across male versus female candidates, we conducted a sequence of increasingly more restrictive tests of invariance across groups (see Hancock, 1997). In particular, the following tests of measurement invariance were conducted: (a) factor form (i.e., the same number of factors and the factors have the same variables that load on them), (b) factor loadings, (c) factor variances and covariances, and (d) errors of measurement.

To determine whether constraining parameters to be invariant across groups yielded a significant decrease in fit, researchers have traditionally

used the $\Delta\chi^2$ as the index of difference in fit. However, the use of $\Delta\chi^2$ has been criticized because of its sensitivity to sample size (Cheung & Rensvold, 2002; Kelloway, 1995). Recently, Cheung and Rensvold (2002) provided evidence that ΔCFI was not prone to these problems. On the basis of extensive simulations, they also determined that a ΔCFI value higher than .010 was indicative of a significant drop in fit. If the ΔCFI indicated that the constrained model did not lead to a significant decrease in fit compared with the unconstrained model, the constrained parameters were considered to be invariant across groups.

*Tests of latent mean differences between men and women.* Finally, we estimated a mean structure in addition to the covariance structure already obtained (i.e., measurement models estimated using confirmatory factor analyses). To this end, we included indicator means in the analysis and examined the latent means for the factors of interest. Specifically, in the ratings of the male candidates, the latent means were freely estimated, whereas in the ratings of the female candidates, they were constrained to be zero (see Hancock, 1997). Next, for the ratings of male candidates, we examined whether the latent means were significantly different from zero. Given statistical significance, there was evidence of a significant latent mean difference between male and female candidates.

*Effect size differences.* Apart from the latent mean analyses, we also computed standardized mean differences (*d* values) between the male and female groups on the observed dimension ratings. Note that positive *d* values indicate that men (the majority group) were rated more favorably by assessors than women; negative *d* values indicate that women (the minority group) were rated more favorably.

## Results

### *Underlying Structure of Assessment Center Ratings*

Prior to examining gender differences at the latent construct level, we first tested several measurement models that represented different conceptualizations of assessment centers. Results of the confirmatory factor analyses per sample are presented in Table 3. In each of the samples, Model 4 was the only model that attained an acceptable fit, with CFI values around .95 and RMSEA values around .06. This model posited that ratings could be best represented by a combination of dimensions and exercises. Furthermore, this model was not plagued by estimation problems (e.g., nonconvergence, improper estimates).

Besides these global fit measures, evidence that the dimensions deserved their place in this assessment center was also ascertained by inspecting the parameter estimates of Model 4. For example, all parameter estimates related to dimensions were significant. On average, dimensions accounted for a large percentage of the variance (an equal 51% in ratings of both male and female candidates), whereas exercises accounted for only about 20% of the variance. In addition, the median intercorrelation among dimensions was around .60 (.60 for men and .61 for women). These values fare better than the mean values (34% for dimension variance and .71 for the dimension factor correlation) computed across a large number of assessment center studies by Lievens and Conway (2001). Inspection of the multitrait–multimethod matrix also attested to the fact that there was construct-related validity evidence for this assessment center. In fact, the average monotrait–heteromethod correlation (indicative of convergent validity) was .53. This is much higher than in other assessment center construct validity studies (e.g., Bycio et al., 1987). The average heterotrait–monomethod correlation (indicative of discriminant validity) was .55.

In sum, the results of these within-group tests of various measurement models indicated that, in this specific assessment center, assessor ratings could be best represented by a combination of both dimensions and exercises. This is an important result because it confirms that dimensions were actually measured in this assessment center. If we had found no evidence for dimension variance and only exercise variance (see, e.g., Bycio et al., 1987), we would not have been able to investigate our hypotheses because these were formulated at the level of the dimensions. The next step then became to examine the invariance of this measurement model across the two groups. In these measurement invariance tests, we concentrated solely on the dimensions (see our hypotheses).

Table 3
*Summary of Goodness-of-Fit Indices for Within-Group Measurement Models*

| Confirmatory factor analysis model | $\chi^2$ | *df* | CFI | RMSEA | CI for RMSEA |
|---|---|---|---|---|---|
| Model 1 | | | | | |
| Correlated dimensions only | | | | | |
| Ratings of male candidates | 6,892.03 | 194 | .654 | .149 | .146–.152 |
| Ratings of female candidates | 1,337.37 | 194 | .635 | .153 | .145–.160 |
| Model 2 | | | | | |
| Correlated exercises only | | | | | |
| Ratings of male candidates | 5,698.48 | 194 | .715 | .135 | .132–.138 |
| Ratings of female candidates | 1,144.80 | 194 | .697 | .139 | .131–.147 |
| Model 3 | | | | | |
| One general dimension and correlated exercises | | | | | |
| Ratings of male candidates | 3,289.03 | 172 | .818 | .108 | .105–.111 |
| Ratings of female candidates | 661.43 | 172 | .824 | .106 | .097–.114 |
| Model 4 | | | | | |
| Correlated dimensions and uncorrelated exercises | | | | | |
| Ratings of male candidates | 1,080.25 | 172 | .961 | .058 | .055–.062 |
| Ratings of female candidates | 351.38 | 172 | .935 | .064 | .054–.074 |

*Note.* $N = 1{,}559$ for ratings of male candidates, and $N = 254$ for ratings of female candidates. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval.

Table 4
*Tests of Measurement Invariance for Multi-Group Measurement Model With Exercises and Dimensions (N = 1.803)*

| Model | $\chi^2$ | *df* | $\Delta\chi^2$ | $\Delta df$ | CFI | ΔCFI | RMSEA |
|---|---|---|---|---|---|---|---|
| Equal number of factors | 1,431.62 | 344 | | | .959 | | .042 |
| Equal factor loadings | 1,470.71 | 376 | 39.09 | 32 | .959 | .000 | .040 |
| Equal factor variances and covariances | 1,532.65 | 403 | 61.94 | 27 | .958 | −.001 | .039 |
| Equal measurement errors | 1,607.93 | 425 | 75.28 | 22 | .956 | −.002 | .039 |

*Note.* CFI = comparative fit index; RMSEA = root-mean-square error of approximation.

### *Tests of Measurement Invariance*

Table 4 presents the results of the sequence of increasingly more restrictive tests of measurement invariance. As mentioned above, the first test was a test of factor form invariance. When we constrained the number of factors and constrained which variables loaded on the factors to be invariant across groups, a good fit was obtained (CFI = .959 and RMSEA = .042). So, we continued with our tests of measurement variance and constrained the factor loadings to be invariant across groups. As fit was still very good (CFI = .959 and RMSEA = .040), we tested for the invariance of factor variances and covariances across groups. The addition of this constraint also did not lead to a significant decrease in fit, as indicated by the fit indices (CFI = .958 and RMSEA = .039). Finally, we tested whether constraining the measurement errors to be invariant across groups still yielded a very good fit. This was indeed the case because constraining the error matrices to be invariant did not detract from fit (CFI = .956 and RMSEA = .039).

Although these results indicated that fit was very good even for the most restrictive model and therefore proved that the measurement model was invariant across groups, some might argue that there was a significant decrease in $\chi^2$ for each of the stacked measurement models (e.g., when factor variances and covariances were constrained, $\chi^2$ dropped from 1,532.65 to 1,470.71), $\Delta\chi^2(27, N = 1{,}803) = 61.94$, $p < .01$, and that therefore some of the constraints imposed on individual parameters were probably not supported. As mentioned in the earlier section, we did not use the $\Delta\chi^2$ as an index for determining whether a constrained model produced a significantly worse fit than an unconstrained model because of its sensitivity to large sample sizes (see Brannick, 1995; Cheung & Rensvold, 2000; Kelloway, 1995, for other drawbacks related to $\Delta\chi^2$). In any case, inspection of modification indices showed that constraints did not hold for only 12 of the 81 constraints imposed on the individual parameter estimates. In addition, the largest $\chi^2$ increase by freeing an individual parameter was only 13.02.

In sum, the results of these multiple-group models of stacked measurement invariance demonstrated that there was no considerable departure from measurement invariance. Hence, it was considered meaningful to compare the (error-free) latent means of the male and female candidates, which was the aim of this study.

### *Test of Latent Means Between Men and Women*

After including indicator means in the analysis, we tested for latent mean differences using the procedure described above.[2] Table 5 shows the results of these latent mean difference tests for constructs and exercises. Given that our hypotheses dealt with gender differences on constructs, we focus here on the constructs. Note, however, that there were also significant latent mean differences on two exercise factors (others command exercise and planning exercise), with men performing better.

As shown in Table 5 (left section), there were significant latent mean differences on communication ($z = -3.45$, $p < .01$), interaction ($z = -4.57$, $p < .01$), and drive and determination ($z = -2.77$, $p < .05$). Female candidates received higher ratings than male candidates on these three constructs. This is consistent with Hypothesis 2, although the significant difference favoring female candidates for drive and determination is unexpected. The latent mean difference tests did not reveal significant gender differences favoring male candidates. This does not lend support to Hypothesis 1.

### *Effect Size Differences on Observed Dimension Ratings*

For comparison purposes with the latent mean analyses, we also computed subgroup (male–female) descriptive statistics and *d* value (Cohen, 1997) differences. Table 5 (right section) reports these statistics computed on the observed overall dimension ratings. As noted above, observed overall dimension scores were obtained by averaging the scores on the same dimension across exercises. The *d*s ranged from −.31 on interaction to .27 on problem solving. A comparison of the results of the latent mean analyses and the *d*s shows that results were generally similar. A notable exception was the dimension of problem solving. Whereas the latent mean analyses did not find a significant difference between men and women, *d* equaled .27.

For comparison reasons with prior research (see Table 1), we also computed *d* associated with the OAR. Female candidates ($M = 2.13$, $SD = 0.98$) were rated notably higher than male candidates on the OAR ($M = 1.92$, $SD = 0.92$, $d = -.22$). Note also that 45.9% ($N = 119$) of the female candidates were accepted, whereas only 31.9% ($N = 507$) of the male candidates were accepted, $\chi^2(1, N = 626) = 19.46$, $p < .001$.

[2] Note that we also tested whether there were latent mean differences by including the multiple-group stacked measurement model without the 12 noninvariant constraints imposed. The results were identical to the results presented. This shows again that it was meaningful to compare the latent mean differences, even if some individual parameter estimates were not invariant across groups.

Table 5
*Differences Between Men and Women on Latent Factors and Observed Dimension Ratings*

| | Latent mean analyses | | Observed mean analyses | | | | |
|---|---|---|---|---|---|---|---|
| | | | Men | | Women | | |
| Dimension–exercise | *z* | *p* | *M* | *SD* | *M* | *SD* | *d* |
| Dimension | | | | | | | |
| Oral communication | −3.45 | < .01 | 8.18 | 1.04 | 8.36 | 1.01 | −.17 |
| Interaction | −4.57 | < .01 | 7.97 | 1.38 | 8.40 | 1.33 | −.31 |
| Impact | −1.43 | *ns* | 8.04 | 2.16 | 7.99 | 2.07 | .02 |
| Problem solving | 0.40 | *ns* | 6.82 | 2.04 | 6.27 | 1.85 | .27 |
| Drive and determination | −2.77 | < .01 | 9.01 | 1.66 | 9.18 | 1.57 | −.10 |
| Resistance to stress | −1.64 | *ns* | 6.19 | 1.14 | 6.26 | 1.02 | −.07 |
| Exercise | | | | | | | |
| Lecturette | −0.12 | *ns* | — | — | — | — | — |
| Other command tasks | 2.38 | < .05 | — | — | — | — | — |
| Own command tasks | 1.64 | *ns* | — | — | — | — | — |
| Opening tasks | 1.63 | *ns* | — | — | — | — | — |
| Group planning exercise | 2.22 | < .05 | — | — | — | — | — |
| Leaderless group discussion | 1.18 | *ns* | — | — | — | — | — |

*Note.* *d* is the difference between male and female means in standard deviation units (effect size). *d* values were computed by expressing the difference between the means of the majority and minority groups in pooled standard deviation units. $d$ = (mean for the majority group − mean for the minority group)/$SD_{pooled}$. Positive *d* values indicate men score higher; negative *d* values indicate that women score higher. Dashes indicate that there were no final exercise ratings made.

## Discussion

This study used a construct-driven approach to shed light onto possible gender differences in assessment centers. Specifically, we extended prior research (a) by examining gender differences at the construct level, (b) by formulating a priori hypotheses about which constructs would be susceptible to gender effects, and (c) by using analytical methods to examine both observed dimension and latent construct differences between male and female candidates. To our knowledge, this study is the first to adopt such a construct-driven approach for examining gender differences and is the first to explore in detail these two complementary analytical approaches.

This study found significant latent mean differences favoring female applicants on the dimensions of oral communication and interaction. These results map well into extant research on gender differences on other selection instruments. In fact, similar results favoring female candidates have been found for situational judgment tests (Hough, Oswald, & Ployhart, 2001), work samples (Schmitt et al., 1996), and employment interviews. Specifically, a construct-driven meta-analysis of Huffcutt et al. (2001) reported a −.13 mean effect size (i.e., favoring women) for interview ratings of applied social skills (e.g., communication skills, interpersonal skills). On a broader level, our results are also consistent with empirical research on gender differences in leadership that showed that women are more effective in interpersonal leadership roles.

From a theoretical point of view, our results partially support role congruity theory (Eagly, 1987; Eagly & Karau, 2002). According to this theory, people generally ascribe more communal characteristics to women and female leaders. Examples of communal characteristics include being helpful, interpersonally sensitive, kind, or participative. Further, role congruity theory posits that women, who include these communal behaviors in their repertoire, are viewed as fulfilling aspects of their female role. Hence, they receive more positive ratings on these communal characteristics. A large body of research has confirmed these propositions of role congruity theory. For example, Davison and Burke (2000) reviewed 49 experimental studies that evaluated male and female candidates whose characteristics had been equated. Results showed that women were preferred over men only in female gender-typed roles. However, these experimental studies suffer from ecological validity limitations, as they typically provide a limited amount of written information of hypothetical male and female applicants to inexperienced raters (Bowen, Swim, & Jacobs, 2000). Therefore, it is important that our study extends the findings related to role congruity theory to a real-life situation of final stage selection into an actual leadership role.

Other results are less in line with role congruity theory. In particular, we did not find gender differences favoring male candidates on impact and problem solving. Role congruity theory would have predicted women to receive lower ratings on these dimensions because these dimensions reflect agentic behaviors that are typically ascribed to men. In addition, our finding of a gender difference favoring women on drive and determination does not support role congruity theory. Drive and determination might be considered more agentic and therefore masculine attributes. Probably, these results deviate from role congruity theory because the female candidates in this assessment center were highly self-selected and motivated to enter the army as officers. Hence, the female candidates of this study might have scored higher on agentic characteristics than the general female population.

The flip side of any field study on gender differences is that it is difficult to disentangle possible rival explanations for the results obtained. This field study is no exception. One explanation for the

higher ratings of female candidates in this study is that these ratings reflect real performance differences. In other words, female candidates actually outperformed male candidates on constructs such as oral communication and interaction. Another explanation is that the ratings are indicative of biases on the part of assessors. This means that our results do not reflect actual differences. Instead, they reveal gender stereotypes that are activated and used when assessors observe and evaluate female candidates in an assessment center for officer entry. The fact that nonrepresentative samples of women participated in this selection process yields a third explanation for the gender differences. As compared with men, there were proportionately fewer women who applied, even though the British Army has actively encouraged female applicants to apply for military jobs. Perhaps some kind of preferential selection (e.g., Heilman & Blader, 2001) occurred along the assessment center process in the sense that assessors gave higher ratings to female candidates so that more women would pass the selection process. Clearly, the field setting of our study precludes us from drawing definitive conclusions about these possible rival explanations. Only laboratory studies can disentangle these various explanations because such studies hold ratee performance levels constant.

Apart from the explanations, the generalizability of our results also needs to be discussed. Generally, we believe that the generalizability of the findings from any assessment center study is primarily a function of the meaning and context of the assessment center dimensions and exercises rather than a function of the target job. The leadership dimensions measured in this particular assessment center tap both cognitive and noncognitive aspects of performance. In addition, the leadership dimensions and their respective definitions (see the Appendix) would not look out of place in many large organizations' selection procedures. This is not surprising, as historically many commercial assessment center designs in both the United Kingdom and the United States have been based on military leadership assessment centers (e.g., Dobson & Williams, 1989; Howard, 1997). With respect to the assessment center exercises used, four of the assessment center exercises (planning exercise, group planning exercise, leaderless group discussion, and lecturette) were not cast in a military context and, therefore, are directly comparable with those commonly found in commercial organization assessment centers. However, the other four exercises (own command tasks, other command tasks, opening tasks, and individual obstacles) were framed in a military context, which might detract from their generalizability to other settings.

The generalizability of our results is also related to the samples used. In this study, the sample of women (263 participants) was much smaller than the sample of men (1,594 participants). In terms of the populations to which our results generalize, we believe our results generalize to male-dominated working populations and for selection into male-dominated job roles. Apart from the army, other examples are top management positions, police jobs, firefighter jobs, and supervisory and management job roles in traditionally male-dominated industries (e.g., construction, engineering, technical industries, etc.). A focus on these populations is interesting from both a research and practical point of view given the dearth of research into these populations (Langan-Fox, 1998).

In terms of methodological implications, this study shows that it is important to distinguish between constructs and methods. It does not make sense to generally state that there do not exist subgroup differences in assessment centers or assessment center exercises. Instead, it is important to look at the constructs they measure. One of the key advantages of a construct-driven approach is that it provides a basis for predicting the validity and subgroup differences of specific constructs measured by other methods in other contexts (Hattrup et al., 1992; Schmitt & Chan, 1998). So, it is valuable to examine whether similar constructs produce majority–minority group differences across other selection methods.

As a second methodological implication, this study used latent mean analyses as a way of better understanding the source of gender differences in assessment centers. As already mentioned above, latent mean analyses have several benefits compared with *t* tests on observed means. A drawback of *t* tests is that when a difference is found, researchers run the risk of misattributing the differences found (Cheung & Rensvold, 2000; Hattrup et al., 1992). For instance, in assessment centers, the observed final dimension ratings are summary ratings of dimensions across exercises (e.g., Goffin et al., 1996). Hence, significant gender differences on observed dimension ratings in assessment centers might reflect differences on the construct being measured. However, these gender differences on observed dimension ratings might also reflect differences on the exercises being evaluated. Finally, the observed gender differences might stem from other sources of variance such as measurement error. The key advantage of latent mean analyses is that they provide researchers with a sophisticated tool for determining whether there exist gender differences at the latent level. This is because in the latent mean analyses, measurement equivalence across groups is a prerequisite prior to examining mean differences across groups. When measurement invariance is accounted for, other sources of variance (differences in the reliability or factorial structure of the ratings) are also accounted for. In addition, when measurement equivalence is established, researchers can test for differences on latent factors of interest. As noted above, in assessment centers, these factors pertain to both exercises and dimensions. So, one is able to test whether there are gender differences on exercises or on dimensions, separating exercise from dimension variance.

These benefits of latent mean analyses for understanding gender differences are well illustrated by the results of this study. For instance, the observed final dimension ratings in Table 5 show that the observed ratings of men are significantly higher than those of women on the dimension of problem solving ($d = .27$). Thus, at first sight, one might conclude that there are gender differences on this dimension. However, the latent mean analyses (see Table 5) do not support this explanation, as there were no gender differences on the latent construct level for problem solving. As noted above, the latent mean analyses did find significant differences in favor of men on two latent exercise factors (planning exercise and other command tasks). As shown in Table 2, problem solving is measured in one of these exercises, namely the other command tasks. So, the gender difference on the observed dimension of problem solving is not really a gender difference on the latent construct of problem solving. Instead, it stems from a gender difference on the exercise (other command tasks). However, on

the basis of the observed means, it is impossible to tell because observed dimension ratings confound dimension and exercise variance. Another example is drive and determination. Although the observed ratings do not show gender differences, the latent mean analyses show a significant difference in favor of women on the latent construct level. All of this demonstrates that reliance on observed means might lead to erroneous conclusions. That is, an observed gender difference might be incorrectly attributed to a gender difference on the dimension being measured, even though it stems from a gender difference on a latent exercise factor. Alternatively, on the basis of observed differences, one might conclude that there is no significant difference, whereas the latent mean analyses indicate the opposite. On a broader level, it should be clear that it is difficult to draw subgroup difference conclusions on the basis of observed dimension ratings. In this particular assessment center and in other assessment centers, the degree of confounding between exercises and dimensions might have a significant impact on the observed dimension ratings. Different assessment centers with different combinations of exercises and dimensions, even if conceptually defined in the same way as in the current assessment center, could well produce different results for those dimensions that are more subject to the confounding problem.

Do these benefits of latent mean analyses for understanding the source of group differences imply that one should examine subgroup differences at the latent construct level? In general, we believe structural equation modeling analyses on the latent constructs and effect size analyses on the observed dimensions are complementary instead of contradictory analytical approaches, as they each serve different purposes. In most instances, practitioners will be primarily interested in differences on observed, rated dimensions and in particular in ascertaining that subgroup differences are minimal and minimized by design considerations in an operational assessment center (see, e.g., Lievens & Klimoski, 2001). However, practitioner and researcher interests overlap to a considerable extent also with regard to investigating differences on latent constructs at the dimension level of analysis. We draw no distinction between the two sets of interests (e.g., N. Anderson, 2005) but rather again argue for the importance of using both analytical methodologies in a construct-driven approach to examining differences beyond the overly simplistic level of the OAR.

Finally, this study has some implications for assessment center practice. A first implication is that it might be worthwhile to go beyond OAR differences to examine differences on the dimensions (see our effect size analyses). If information about which dimensions favor female candidates is available, we suggest adjusting how these dimensions are weighted to form an OAR. Similar implications for weighting schemes have been presented in the case of race differences (Pulakos & Schmitt, 1996). A second practical implication concerns the selection of assessment center exercises. Clearly, job relatedness should remain the key criterion for selecting and designing assessment center exercises. If one is also concerned about possible subgroup differences, we suggest combining job-related exercises that primarily tap interpersonally oriented constructs with job-related exercises that capture cognitively oriented constructs in one assessment center. Accordingly, subgroup gender differences will be balanced out. Similarly, one might include job-related exercises measuring interpersonally oriented constructs to balance out subgroup differences favoring men on specific cognitive ability constructs (e.g., test of numerical ability).

Taken together, the present construct-driven study found notable differences on several leadership-oriented dimensions that would not have become evident from examination of OAR score differences alone. For practitioners and researchers alike, such differences are of import in all selection situations but particularly so for final-stage assessment centers into a male-dominated job role. Our findings suggest that it is imperative for future studies to investigate differences at the construct-level, rather than just at the level of the OAR.

## References


Alexander, H. S., Buck, J. A., & McCarthy, R. J. (1975). Usefulness of the assessment center process for selection to upward mobility programs. *Human Resource Management, 75,* 11–13.

Anderson, L. R., & Thacker, J. (1985). Self-monitoring and sex as related to assessment center ratings and job performance. *Basic and Applied Social Psychology, 6,* 345–361.

Anderson, N. (2005). Relationships between practice and research in personnel selection: Does the left hand know what the right is doing? In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 1–24). Malden, MA: Blackwell Publishing.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154.

Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55,* 985–1008.

Baron, H., & Janman, K. (1996). Fairness in the assessment centre. *International Review of Industrial and Organizational Psychology, 11,* 61–114.

Bentler, P. M. (1995). *EQS: Structural equations program manual* [Computer software manual]. Encino, CA: Multivariate Software.

Bobrow, W., & Leonards, J. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality, 12,* 217–236.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67,* 3–19.

Bowen, C., Swim, J. K., & Jacobs, R. R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology, 30,* 2194–2215.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201–213.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230–258.

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.

Champion, C. H., Green, S. B., & Sauser, W. I. (1988). Development and evaluation of shortcut-derived Behaviorally Anchored Rating Scales. *Educational and Psychological Measurement, 48,* 29–41.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross Cultural Psychology, 31,* 187–212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Cleveland, J. N., Stockdale, M., & Murphy, K. R. (2000). *Women and men in organizations: Sex and gender issues at work.* Mahwah, NJ: Erlbaum.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

Collins, J. M., & Gleaves, D. H. (1998). Race, job applicants, and the Five-Factor Model of Personality: Implications for Black psychology, industrial/organizational psychology, and the Five-Factor Theory. *Journal of Applied Psychology, 83,* 531–544.

Davison, H. K., & Burke, M. J. (2000). A meta-analysis of sex discrimination in simulated selection contexts. *Journal of Vocational Behavior, 56,* 225–248.

Dobson, P., & Williams, A. (1989). The validation of the selection of male British Army officers. *Journal of Occupational Psychology, 62,* 313–325.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NJ: Erlbaum.

Eagly, A. H., Johannesen-Schmidt, M. C., & Van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing men and women. *Psychological Bulletin, 129,* 569–591.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin, 108,* 233–256.

Eagly, A. H., & Karau, S. J. (1991). Gender and the emergence of leaders: A meta-analysis. *Journal of Personality and Social Psychology, 60,* 685–710.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109,* 573–598.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin, 117,* 125–145.

Eagly, A. H., Karau, S. J., Miner, J. B., & Johnson, B. T. (1994). Gender and motivation to manage in hierarchic organizations: A meta-analysis. *Leadership Quarterly, 5,* 135–159.

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.

Goldstein, H. W., Yusko, K. P., Braverman, E. P., Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51,* 357–374.

Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). Exploring Black–White subgroup differences of managerial competencies. *Personnel Psychology, 54,* 783–807.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30,* 91–105.

Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology, 77,* 298–308.

Heilman, M. E., & Blader, S. L. (2001). Assuming preferential selection when the admissions policy is unknown: The effects of gender rarity. *Journal of Applied Psychology, 86,* 188–193.

Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50,* 455–470.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13–52.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology, 62,* 429–440.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897–913.

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology, 52,* 536–546.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16,* 215–224.

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 and Personality, 12,* 129–144.

Langan-Fox, J. (1998). Women's careers and occupational stress. *International Review of Industrial and Organizational Psychology, 13,* 273–304.

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., & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? *International Review of Industrial and Organizational Psychology, 16,* 246–286.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management, 20,* 439–464.

Moses, J. L. (1973). The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology, 26,* 569–580.

Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology, 60,* 527–529.

Ployhart, R. P., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7,* 27–65.

Powell, G. N. (1993). *Women and men in management* (2nd ed.). Thousand Oaks, CA: Sage.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68,* 227–231.

Robertson, I. T., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60,* 187–195.

Ryan, A. M., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52,* 37–58.

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

Schmitt, N. (1993). Group composition, gender, and race effects on assessment center ratings. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 315–332). Hillsdale, NJ: Erlbaum.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousands Oaks, CA: Sage.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. *International Review of Industrial and Organizational Psychology, 11,* 115–139.

Schmitt, N., & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology, 62,* 261–264.

Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86,* 451–458.

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. (1992). Subtle gender bias in the assessment of managerial potential. *Sex Roles, 27,* 499–515.

Shore, T. H., Tashchian, A., & Adams, J. S. (1997). The role of gender in a developmental assessment center. *Journal of Social Behavior and Personality, 12,* 191–203.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397–423.

Vernon, P. E., & Parry, J. B. (1949). *Personnel selection in the British forces.* London: University of London Press.

Walsh, J. P., Weinberg, R. M., & Fairfield, M. L. (1987). The effects of gender on assessment centre evaluations. *Journal of Occupational Psychology, 60,* 305–309.

Weijerman, E. A. P., & Born, M. P. (1995). De relatie tussen sekse en assessment center beoordelingen [The relationship between gender and assessment center scores]. *Gedrag en Organisatie, 8,* 284–292.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29,* 231–258.

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8,* 259–296.

## Appendix

### Description of Dimensions

| Dimension | Definition |
|---|---|
| Oral communication | The ability to relate fluently to others, to convey clear meaning without hesitancy and without using slang or dialect. |
| Interaction | The ability of the candidate to relate to others individually and within groups. |
| Problem solving | The ability of the candidate to deal with concrete, factual problems in a practical manner, exercising common sense and judgment. |
| Impact | The extent to which a candidate naturally assumes the lead of the group, influences others, maintains control, and displays initiative. |
| Drive and determination | The extent to which a candidate maintains a consistently high level of activity, does not give up, resolves difficulties, and overcomes natural fears. |
| Reaction to stress | The ability of the candidate to function effectively when under a degree of stress. |
| Analysis and planning[a] | The ability to understand primary and secondary aims, and the use of logic and reasoning to formulate a workable plan. |
| Physical ability[a] | The extent to which a candidate possesses the agility, robustness, and coordination to carry out the physical tasks inherent in the duties of a young officer. |

[a] These dimensions were not included in our analyses (see Method section).

Received January 20, 2004
Revision received April 11, 2005
Accepted April 29, 2005 ■