Received: 29 June 2018 | Revised: 30 December 2018 | Accepted: 18 January 2019

DOI: 10.1111/ijsa.12229

FEATURE ARTICLE

WILEY International Journal of SELECTION AND ASSESSMENT

Same-gender and same-race bias in assessment center ratings: A rating error approach to understanding subgroup differences

George C. Thornton III[1] | Deborah E. Rupp[2] | Alyssa M. Gibbons[1] | Adam J. Vanhove[3]

[1]Department of Psychology, Colorado State University, Fort Collins, Colorado

[2]Department of Psychological Sciences, Purdue University, West Lafayette, Indiana

[3]School of Strategic Leadership Studies, James Madison University, Harrisonburg, Virginia

Correspondence
Adam J. Vanhove, School of Strategic Leadership Studies, James Madison University, Lakeview Hall MSC 1505, 298 Port Republic Road, Harrisonburg, VA 22807.
Email: vanhovaj@jmu.edu

Abstract

This study investigated leniency and similar-to-me bias as mechanisms underlying demographic subgroup differences among assessees in assessors' initial dimension ratings from three assessment center (AC) simulation exercises used as part of high-stakes promotional testing. It examined whether even small individual-level effects can accumulate (i.e., "trickle-up") to produce larger subgroup-level differences. Individual-level analyses were conducted using cross-classified multilevel modeling and conducted separately for each exercise. Results demonstrated weak evidence of leniency toward White assessees and similar-to-me bias among non-White assessee–assessor pairs. Similar leniency was found toward female assessees, but no statistically significant effects were found for assessee or assessor gender or assessee–assessor gender similarity. Using traditional *d* effect size estimates, weak individual level assessee effects translated into small but consistent subgroup differences favoring White and female assessees. Generally small but less consistent subgroup differences indicated that non-White and male assessors gave higher ratings. Moreover, analyses of overall promotion decisions indicate the absence of adverse impact. Findings from this AC provide some support for the "trickle-up" effect, but the effect on subgroup differences is trivial. The results counter recent reviews of AC studies suggesting larger than previously assumed subgroup differences. Consequently, the findings demonstrate the importance of following established best practices when developing and implementing the AC method for selection purposes to minimize subgroup differences.

KEYWORDS
adverse impact, assessment centers, rating error

1 | INTRODUCTION

The assessment center (AC) method has long been viewed as relatively free of adverse impact (Thornton & Rupp, 2006). This position was first established by Huck and Bray (1976) when examining Black–White differences among females in a management AC and has been widely maintained in the AC literature on gender and race subgroup differences for nearly 40 years (Thornton & Byham, 1982; Thornton & Rupp, 2006). ACs often demonstrate strong criterion-related validity (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987), sometimes even exceeding that of cognitive ability tests (Sackett, Shewach, & Keiser, 2017). ACs can also produce incremental predictive validity beyond situational judgment tests (Blair, Hoffman, & Ladd, 2016) and other alternative selection measures (Meriac, Hoffman, Woehr, & Fleisher, 2008; Thornton, Rupp, & Hoffman, 2015). Thus, the AC method has been widely accepted as a valuable selection tool for organizations seeking to maximize both prediction and fairness and to avoid legal challenge (Hoffman & Thornton, 1997).

Multiple recent reviews, however, have drawn attention to the variability in subgroup difference estimates observed across AC studies (d = 0.03–0.60; Bobko & Roth, 2013; Ployhart & Holtz, 2008), and a recent meta-analysis found that at least the Black–White assessee subgroup difference appears to be larger than previously believed (d = 0.52; Dean, Roth, & Bobko, 2008). Given these findings, we cannot simply assume the absence of subgroup differences or adverse impact in AC ratings, particularly for employment decision-making purposes. Research is needed to explore the potential mechanisms that foster subgroup-level differences in AC ratings.

The purpose of the current study is to examine two kinds of individual-level assessor errors, namely leniency and similar-to-me bias, as contributing to subgroup-level differences. We propose that small but systematic biases demonstrated by assessors toward assessees of different demographic subgroups can accumulate or "trickle-up" to produce larger subgroup-level differences. Larger subgroup-level differences, in turn, increase the potential for observing evidence of adverse impact.

Examining individual-level assessor leniency and similar-to-me bias as drivers of subgroup-level differences in ACs makes at least two unique contributions to the literature. First, existing evidence has focused almost exclusively on differences in ratings *received* by different groups of assessees (e.g., Black-White), with little-to-no attention given to differences in ratings *given* by different groups of assessors or as a result of different *assessee–assessor* combinations. Second, existing evidence has relied on forms of aggregated ratings, whether they be post-exercise dimension ratings (PEDRs), which represent an integration of multiple assessors' ratings within an AC simulation exercise, or overall assessment ratings (OARs), which further integrate PEDRs across exercises. Existing evidence comparing aggregated ratings of different demographic subgroups of assessees speaks directly to the level of fairness and legal defensibility associated with the use of ACs. However, PEDRs and OARs lack the richness of information needed to explain *threats* to AC fairness and legal defensibility, when subgroup differences are observed. PEDRs and OARs represent an aggregation of many different ratings of assessees, often by many different assessors. Deconstructing these aggregated data into individual-level observations provides one viable avenue for understanding how subgroup-level differences accumulate.

We test for individual-level assessor biases using ratings from three AC exercises used as part of a high-stakes promotional testing program in a police department located in a major U.S. metropolitan area. Assessee and assessor main effects (leniency) and assessor-by-assessee interaction effects (similar-to-me bias) are tested using individual assessors', pre-consensus ratings. These effects are tested separately for the demographic characteristics of race and gender using the cross-classified multilevel modeling technique (Fielding & Goldstein, 2006; Hox, 2002). This procedure overcomes violations of the independence of observation assumption inherent to analysis of variance techniques commonly used in the AC subgroup difference literature. The procedure does so by allowing for assessee and assessor variance components to be disentangled, while simultaneously accounting for the dependencies among ratings of the same assessee by different assessors and among ratings of different assessees by the same assessor. Thus, the procedure is consistent with our purpose of examining the systematic presence of rater error in individual-level, pre-consensus ratings observed among assessees, assessors, and assessee–assessor combinations as a mechanism driving subgroup-level differences. We further examine the extent to which these potential systematic biases "trickle-up" into race and gender subgroup-level differences in PEDRs and influence conclusions regarding the presence of adverse impact in final promotion decisions.

2 | SUBGROUP DIFFERENCES IN ASSESSMENT CENTER RATINGS

2.1 | Assessee subgroup-level differences

Substantial evidence exists regarding assessee race and gender subgroup differences. Early studies of subgroup differences in ACs found minimal evidence of assessee differences due to race (Huck & Bray, 1976) or gender (Moses & Boehm, 1975), and subgroup differences that were found were typically thought to be smaller than those observed in paper-and-pencil tests (Schmitt & Mills, 2001). More recent reviews, however, have cited Black–White assessee subgroup differences in primary studies to be as large as d = 0.60 among exercises and dimensions that are strongly cognitive in nature (Bobko & Roth, 2013; Ployhart & Holtz, 2008). Further, a meta-analysis by Dean et al. (2008) shows a substantial Black–White difference among assessee ratings (d = 0.52), suggesting that White assessees are rated, on average, approximately one-half of a standard deviation (SD) higher than Black assessees. Findings also show a Hispanic-White difference (d = 0.28) favoring White assessees and a gender difference (d = 0.19) favoring females. Evidence that subgroup difference estimates can vary considerably across studies (Bobko & Roth, 2013; Ployhart & Holtz, 2008), as well as larger Black–White mean differences than had previously been assumed (Dean et al., 2008), raise considerable concern over the fairness and legal defensibility of ACs, suggesting the need to identify explanatory mechanisms driving these differences.

2.2 | Assessor and assessor–assessee subgroup-level differences

Assessees are not the only possible source of subgroup differences in AC ratings. Assessors are also diverse, and this creates the potential for rating differences across groups of assessors and across combinations of assessors and assessees. We identified only four studies comparing ratings among assessor and/or assessee–assessor race or gender subgroups. Only one study has examined subgroup differences due to assessor and assessor–assessee race. Schmitt (1993) found trivial mean level differences in ratings given by Black and White assessors, overall, but did find that Black assessors gave higher ratings to Black assessees than did White assessors.

Three studies have examined assessor gender subgroup differences, with two showing no subgroup differences in ratings made by male and female assessors (Schmitt, 1993; Walsh, Weinberg, & Fairfield, 1987). Shore, Tashchian, and Adams (1997) found no support for assessor gender effects in two AC exercises but did find that female assessors rated assessees higher in a third exercise.

Research on subgroup differences among assessee–assessor gender combinations shows more mixed results. Falk and Fox (2014) and Shore et al. (1997) found no evidence of subgroup differences due to an assessee–assessor gender interaction. Schmitt (1993) found statistically significant gender interaction effects for multiple AC dimensions, but concluded that these constituted only a small fraction of the many interactions tested and that the practical significance of these interactive gender effects was small. In an AC with predominantly male assessees and assessors, Walsh et al. (1987) found a statistically and practically significant gender interaction effect, in which all-male assessor panels rated female assessees higher than male assessees.

The lack of evidence regarding assessor and assessor-by-assessee subgroup differences may be due to the fact that assessors' race and gender often go unreported (Fiedler, 2001; Goldstein, Yusko, Braverman, Smith, & Chung, 1998). In fact, none of the 27 studies comprising Dean et al.'s (2008) meta-analysis reported assessor demographics. Moreover, even when assessors' race and gender are available for study, AC data are often in the form of post-consensus PEDRs, which represent aggregated assessor ratings. Thus, analyses of subgroup differences involving assessor demographic variables (i.e., assessor and assessor–assessee subgroup comparisons) can only be conducted based on the demographic composition of assessor panels (e.g., all male, all female, mixed). This was the case in two of the studies reviewed above (Schmitt, 1993; Walsh et al., 1987). A third by Falk and Fox (2014) examined data across various offerings of an ongoing AC used to select low-level managers within a government organization. Assessor subgroup differences reported in the study were even further removed from actual assessor effects than assessor panel compositions, relying instead on the overall proportion of male and female assessors employed in specific offerings of the AC to compare differences in mean ratings between offerings. Of the four studies reviewed above, only Shore et al. (1997) estimated non-aggregated assessor effects, using only one assessor per assessee, per exercise.

The aggregation of individual assessor ratings can potentially be beneficial, as it may minimize the impact of one or more individual assessors' biases. This is most likely to be the case when individual assessors' ratings are aggregated through consensus discussion, where assessors' biases may be corrected when assessors must explicitly justify their ratings with behavioral observations. This technique can also fail when influential biased assessors dominate the discussion. However, of greater concern is when individual assessor ratings are aggregated statistically. This is an increasingly common practice (Thornton et al., 2015) and can bias assessor post-consensus ratings and, in turn, assessee subgroup differences to the extent that individual errors are exacerbated, as opposed to minimized, through aggregation. Research has yet to explore the effect of individual-level pre-consensus ratings, and associated rating error, on post-consensus PEDRs and OARs used to estimate assessee subgroup differences. However, research has demonstrated that even small subgroup differences in employee promotion rates at lower levels can produce large disparity at higher levels of the organization (Martell, Lane, & Emrich, 1996). Likewise, even minimal levels of rating error committed by individual assessors, when systematic, may exacerbate subgroup differences and the potential for the presence of adverse impact. In this study, we explore whether demographic differences among assessees and assessors are associated with particular rating errors at the individual assessor level, and whether, in turn, these errors "trickle up" to create subgroup differences in outcomes.

3 | RATER ERROR AS A MECHANISM DRIVING SUBGROUP DIFFERENCES

Because subgroup-level differences in evaluations due to assessee and assessor demographic characteristics are important from ethical, practical, and legal perspectives, it is important to understand their origins. One potential explanation is that these differences may be the result of systematic individual-level assessor errors. Rater errors are thought to be due, in part, to limitations of raters' information processing capabilities regarding various aspects of the evaluation process, including: encoding, categorizing, and integrating observed behaviors into accurate summary judgments of performance (Kane, 2000; Murphy & Cleveland, 1995). The inevitable presence and possible consequences of rater error are well-known throughout the broader subjective performance rating literature (Borman, 1977; Cooper, 1981; Landy & Farr, 1980). While best practice recommendations for minimizing sources of rater bias were first established more than 40 years ago (International Task Force, 2015), a range of issues can arise throughout AC design and execution phases that increase rating error. Examples include failing to adequately train assessors, not including post-exercise assessor discussions to eliminate individual rating biases and errors from post-consensus ratings, asking assessors to work long hours, and not giving assessors sufficient time to evaluate and score assessees (Caldwell, Thornton, & Gruys, 2003; Dewberry, 2011; Dewberry & Jackson, 2016).

Despite assumptions about the presence of rater error and the possible conditions under which it is exacerbated, empirical evidence capturing the effects of such errors on the magnitude of subgroup differences is strangely absent from the AC literature. Perhaps this is because directly measuring assessors' individual biases toward social groups presents a substantial methodological challenge (Greenwald, McGhee, & Schwartz, 1998). However, we propose that it is at least possible to test the degree to which individual assessors demonstrate patterns of ratings consistent with specific types of rating error, and whether commission of these errors varies systematically as a function of the demographic characteristics of assessors or assessees.

3.1 | Assessee and assessor main effects: Leniency error

With the vast majority of existing AC research focusing on assessee subgroup differences, there is an implied focus on the systematic accumulation of *leniency* demonstrated across individual-level ratings of assessees representing particular demographic categories. More specifically, evidence of assessee subgroup differences favoring White and female assessees (Dean et al., 2008) may suggest that White and female assessees tend to receive systematically more lenient ratings from individual assessors, and thus are favored in employment decisions made as a result of AC evaluations.

Leniency errors based on demographic characteristics, such as race and gender are consistent with longstanding theories of stereotypes (e.g., Allport, 1954; Katz & Braly, 1933), which argue that people ascribe both positive and negative characteristics to individuals on the basis of their membership in identifiable groups. Stereotypes associated with being White are largely positive (Fiske, Cuddy, Glick, & Xu, 2002), so an assessor evaluating a White assessee might assume or infer positive characteristics that the assessee has not directly displayed, thus inflating that assessor's rating of that candidate.

Stereotypes associated with being female are somewhat more complex. Despite consistent findings that women are underrepresented in many professions and are frequent targets of many kinds of discrimination in the workplace (Koenig, Eagly, Mitchell, & Ristikari, 2011; Neumark, Bank, & Van Nort, 1996), when people are asked directly about their stereotypes of women in general, these stereotypes are often highly positive (Eagly & Mladinic, 1989; Glick et al., 2004). Heilman (1983) noted that stereotypes of women include positive characteristics such as warmth and sensitivity as well as negative characteristics such as dependence or passivity: a profile which may be seen as incompatible with leadership roles (cf. Eagly & Karau, 2002). However, Fiske and colleagues (2002) found that "business women" in particular were rated as highly competent but only moderately warm. Although these findings may appear counterintuitive, in light of overall patterns that suggest advantages for men in the workplace (cf. Bowen, Swim, & Jacobs, 2000), they are consistent with empirical findings suggesting that female assessees receive slightly higher ratings in ACs (e.g., Dean et al., 2008). Thus, we might also expect that an assessor who endorses these stereotypes might ascribe positive characteristics to a female assessee, inflating ratings on some if not all dimensions.

Stereotypes are often widely shared, with many members of a society ascribing the same characteristics to the same groups of people (e.g., Fiske et al., 2002; Katz & Braly, 1933). Indeed, even those who are members of a particular group often endorse widespread stereotypes about that group (e.g., Tinsley, Howell, & Amanatullah, 2015). Thus, if many assessors share a stereotype about a particular group, and each of those assessors' ratings is slightly inflated for assessees who belong to that group, the aggregated ratings for members of that group may be sufficiently inflated to produce a noticeable subgroup-level difference. This would appear in ratings as an *assessee main effect*, a consistent difference in assessee ratings based on the assessee's race or gender, regardless of the characteristics of the assessor. We therefore tested the following hypotheses regarding main effects of assessee race and gender on leniency error:

> **Hypothesis 1a** *Ratings of White assessees will exhibit greater leniency error than ratings of non-White assessees.*

> **Hypothesis 1b** *Ratings of female assessees will exhibit greater leniency error than ratings of male assessees.*

Systematic differences in leniency error associated with *assessor* characteristics are also of interest, though there is less theoretical basis to understand such effects. Existing empirical evidence regarding assessor subgroup differences in ACs generally suggests statistically and practically non-significant effects. However, this body of evidence is severely limited in quantity—with one study examining assessor race (Schmitt, 1993) and three studies examining assessor gender effects (Schmitt, 1993; Shore et al., 1997; Walsh et al., 1987)—and by the fact that assessor effects are often operationalized through assessor panel composition, rather than by studying the ratings of individual assessors, as discussed above.

However, evidence outside of the AC literature has demonstrated non-aggregated assessor race effects. For example, performance appraisal evidence collected from approximately 40,000 military personnel showed Black and Hispanic raters tended to give higher performance ratings than White raters (Pulakos, White, Oppler, & Borman, 1989), a field study of supervisor-subordinate dyads found that older raters gave higher ratings than younger raters (Griffeth & Bedeian, 1989), and a meta-analysis of gender bias in performance appraisals found that (pro-male) bias was more likely when all of the raters involved in the appraisal were male (Bowen et al., 2000). The authors of the latter study explained this finding in terms of in-group favoritism, which we will revisit below, but they also note that in their study as well as in two prior meta-analyses, there was no comparable effect found for all-female rater groups (i.e., only male raters appeared to demonstrate this particular bias). These findings suggest the possibility of *assessor main effects*—that is, that there may be consistent tendencies for assessors who share demographic characteristics to rate in similar ways.

Explanations for rater main effects are typically scant, although Griffeth and Bedeian (1989) suggest a promising theoretical rationale when they discuss the main effect of rater age in terms of career stages and cohorts. They argue that raters with similar experiences are likely to have more similar views of the requirements of the job, which may affect their frame of reference for making ratings. This is consistent with work by Landy and Vasey (1991), who found that police officers with different demographic characteristics described their jobs differently in a job analysis task; as well as findings within the AC literature that assessors with different training backgrounds (human resources vs. industrial-organizational psychology) rate in predictably different ways (Thornton & Rupp, 2006).

Before we can understand the drivers of assessor main effects, however, we must establish the existence and magnitude of such effects. Our use of individual-level ratings, made independently by assessors before consensus discussion or statistical aggregation, allows us to assess the degree to which individual assessors vary in their tendency to commit leniency error and, if so, whether any part of this variance can be explained systematically by assessor race or gender. We examined the following hypotheses regarding differences in systematic leniency error as a function of assessor race and gender characteristics:

Hypothesis 1c *Ratings made by non-White assessors will exhibit greater leniency error than White assessors.*

Hypothesis 1d *Ratings made by female assessors will exhibit greater leniency error than male assessors.*

3.2 | Assessee–assessor interaction effects: Similar-to-me bias

The similarity-attraction paradigm (Byrne, 1971) is particularly relevant in describing how individual-level ratings made by assessors of assessees of the same race or gender can systematically accumulate to produce subgroup-level differences in ACs. The effects of similarity on attraction, liking, and positive expectations and evaluations have been supported across a wide range of dyads, including: leaders and followers (Sears & Holmvall, 2010), job applicants and organizations (Schreurs, Druart, Proost, & DeWitte, 2009), and trainers and trainees (Varela, Cater, & Michel, 2011). In an AC context, similarity of an assessor to an assessee may affect several stages of the observation and evaluation process inherent to the method (Thornton et al., 2015): assessors' selective expectations for assessee performance, selective attention to confirming evidence of positive behaviors demonstrated by assessees, and the assignment of more favorable judgments to assessees.

As such, if certain demographic groups are not represented within the *assessor* pool, we should expect that this would lead to biased ratings across demographic groups within the *assessee* pool. This assumption is partially responsible for recommendations made within professional guidelines to have a diverse assessor panel (International Task Force, 2015). However, there is little empirical evidence to establish the prevalence or extent of similar-to-me bias in actual ACs. The only available study (Schmitt, 1993) found a same-race effect, with candidates who were of the same race as the majority of the assessor panel receiving higher ratings. Evidence from the performance appraisal literature using non-aggregated rater data corroborates these findings, with meta-analytic (Kraiger & Ford, 1985) and large-scale field study (Pulakos et al., 1989) evidence suggesting same-race bias among White, Hispanic, and Black ratee-rater combinations. Thus, using independent, pre-consensus assessor ratings, we propose the following hypothesis:

Hypothesis 2a *Ratings by White assessors of White assessees will be higher than ratings of non-White assessees. Ratings by non-White assessors of non-White assessees will be higher than ratings of White assessees.*

As is the case with stereotypes, similarity effects appear to be more complex with regard to gender than with regard to race. In the previously mentioned meta-analysis of performance appraisals by Bowen and colleagues (2000), studies that included only male raters showed pro-male bias, but studies that included both male and female raters did not show significant bias. However, in the only study to test interactions of assessor and assessee gender in an AC context, all-male assessor panels gave females higher ratings (Walsh et al., 1987). Further, this interaction is consistent with findings from a large online opinion survey (Elsesser & Lever, 2011) that male employees tended to rate female managers more positively and female employees tended to rate male managers more positively. Elsesser and Lever proposed that this difference might be explained by differences in perceived competition with the opposite gender (consistent with theories of intergroup conflict; e.g., Sherif, 1966; Tajfel, 1978). They found that participants were more likely to perceive competition from the same gender than the opposite gender, and that those participants who did perceive such competition rated same-gender managers less positively. Because the AC in the current study involves predominantly male assessors and assessees and is set in a traditionally male-dominated occupation, we expect that perceptions of same-gender competition are likely, creating conditions under which cross-gender preferences are may occur. We therefore propose the following hypothesis with regard to pre-consensus individual-level assessor ratings:

Hypothesis 2b *Ratings by male assessors of female assessees will be higher than ratings of male assessees.*

3.3 | Ancillary investigations: Dimensions and exercises

The inclusion of multiple dimensions and multiple exercises are essential elements of the AC method. Assessors make observations and judgments of behaviors relevant to dimensions as elicited by exercises. Research has demonstrated the differing nature of dimensions and simulation exercises, the role of cognitive ability in performance effectiveness, and racial subgroup differences (Bobko & Roth, 2013; Buckett, Becker, & Roodt, 2017; Ployhart & Holtz, 2008; Roth, Bobko, McFarland, & Buster, 2008). This raises the question of whether assessor errors occur in different ways in different dimensions and different exercises, and whether they are influenced by assessor and assessee demographic variables. The present study provided an initial opportunity to examine whether rating errors may be operating differently based on the type of dimension and simulation exercise. Based on the limited research in this area, we posed the following research questions:

Research Question A: Do rating errors differ across types of exercise?

Research Question B: Do rating errors differ across types of dimensions?

4 | METHOD

4.1 | Sample

Assessees consisted of 198 police officers who had passed the written examination phase of a promotional process and were vying for promotions to sergeant, for which there were 50 openings. Seventy-seven of the assessees were White, 81 were Black, 36 were Hispanic, and four were not classified by race. The assessee pool was 89.3% male. Assessors consisted of 49 officers holding the rank of sergeant or above from comparable cities throughout the United States. Twenty of the assessors were White, 17 were Black, 10 were Hispanic, and 2 were not classified by race. The assessors were 80.9% male. Although race and gender were not balanced within assessee and assessor samples, they were both representative of the workforce in this context. More importantly, the composition of the pool of assessors matched the composition of the assessee pool. Due to the legal context within this employment environment, the organization followed a strict policy of *not* considering demographic characteristics in any way in the administration of the assessment and decision-making process, and thus it assigned assessors randomly to observe assessees. Analyses of race effects included comparisons of only White and non-White groups, due to the small number of Hispanic assessees and assessors within this sample.

4.2 | Assessment center method

The assessment center that generated the data used in this study was one component of a police sergeant promotional process. The AC included three exercises through which six different performance dimensions (derived via job analysis) were evaluated: Written Communication, Oral Communication, Customer Service, Problem Solving/Decision Making, Conflict Resolution/Teamwork, and Leadership. The Tactical Analysis exercise required candidates to describe how they would handle a complex crisis intervention, in terms of analyzing the situation and deciding what action to take. The In-box exercise had candidates read a complex set of incoming letters, memos, and phone messages; make decisions; and write replies and directives. In the Oral Presentation exercise, candidates were required to describe how they would discuss with subordinates an interpersonal problem in the work unit. All six dimensions were rated in each simulation.

Assessees were required to attend an AC orientation, which covered the dimensions and exercises, along with tips on how best to approach the process. Assessees' overall AC scores were combined arithmetically with their scores on the written test of knowledge of rules and regulations for police operations to yield a final promotional score used to make promotion decisions.

Assessor training conformed to the AC *Guidelines* (International Task Force, 2015). Assessors were sent preliminary materials including information about the police department and job descriptions. On-site training lasted two days and consisted of descriptions of the dimensions to be assessed, practice with the exercises, review of potential errors of observation and rating, practice and feedback on rating behavior, and opportunities to practice interacting with candidates. The competence of each assessor to carry out his or her duties was verified at the end of training.

Assessors and assessees were asked to examine a list of their counterparts and indicate any personal or business relationships, in which case the assessor was prevented from assessing that assessee (that is, a different assessor would be assigned). With these exceptions, for each exercise, assessees were assigned randomly to a two-person assessor panel. If an assessee was assigned to an assessor for more than one exercise, adjustments were made to ensure that each assessee was observed by six different assessors.

Other than three standardized questions asked by assessors at the end of each simulation exercise, no interaction was allowed between assessees and assessors. Assessors observed assessee behavior and took written notes. Behaviorally anchored rating scales were provided to guide assessors' observations and ratings. After an assessee left the examination room, the assessors independently (i.e., without conferring with each other) rated each dimension within that exercise on a scale from 1 to 5, using 0.5 intervals (e.g., 3.5). If the two assessors differed by more than one point on any dimension in their initial independent ratings, they compared observations of behaviors, and were required to come to consensus within one point. No further discussion was allowed. All subsequent integration of ratings was done arithmetically to yield scores on dimensions (the average of two assessors).

An overall assessment score (OAR) was calculated by averaging across dimensions and exercises. The OARs were standardized and weighted (55%), then combined with the standardized and weighted knowledge test scores (45%) to yield the final promotional exam scores. The final promotional exam scores were used to make promotion decisions on a strict top-down basis. For this specific promotional exam, there were 50 openings, meaning the 50 highest scoring assessees would be offered promotion.

This promotional process is quite similar to the long-standing practices of other city, county, and state police and fire departments (Barrett, Doverspike, & Young, 2010; Coleman, 1992; Thornton & Byham, 1982). Agencies often use a knowledge test, either as an initial screening step or in combination with AC scores derived from a procedure using dimensions, exercises, and the types of rating procedures employed here to yield overall scores to inform promotion decisions.

4.3 | Analytic strategy

The present study includes three sets of analyses. Each analysis represents an incremental level of testing for the presence of discrimination in this high-stakes promotional AC. First, at the *individual* level, we used cross-classified multilevel modeling to test for statistical significance in the differential presence of leniency and similar-to-me bias among assessors' independent, pre-consensus dimension ratings due to assessee, assessor, and assessee–assessor race and gender. Second, we examined whether these individual-level effects translated into *subgroup-level differences* among aggregated dimension ratings. Subgroup differences were assessed using traditional mean difference *effect size estimates* (see Ployhart & Holtz, 2008). Third, we examined whether actual promotion decisions, based in part on AC OARs, resulted in the presence of *adverse impact* among one or more demographic groups.

Descriptive statistics for data used across analyses are presented in Tables 1 and 2. Means and standard deviations for ratings received by race and gender subgroups of assessees are presented in Table 1, and for ratings given by race and gender subgroups of assessors in Table 2. Tables 3 and 4 provide means of ratings for race-similar and gender-similar pairs of assessors and assessees.

4.4 | Individual-level ratings: Cross-classified multilevel modeling

Much of the prior research on ratee and rater main and interactive effects has relied on an analysis of variance (ANOVA) framework (e.g., Lin, Dobbins, & Farh, 1992; McFarland, Ryan, Sacco, & Kriska, 2004; Prewett-Livingston, Field, Veres, & Lewis, 1996). However, ANOVA does not appropriately account for the complex structure of typical AC data. In most ACs, each assessor rates multiple assessees, so ratings of different assessees by the same assessor are not statistically independent (i.e., assessees are nested within assessors). In the AC reported on here, each pair of assessors followed the procedure of making independent ratings without consulting with each other. However, since the same assessor rates all dimensions within an exercise, the dimension ratings cannot be said to be independent in the statistical sense. Also, each assessee was rated by two assessors in each exercise, so ratings of the same assessee by different assessors are not statistically independent as they are based on observations of the same behaviors. Further, assessors were not consistently paired into teams. That is, for example, Assessor A might rate the same assessees as Assessor B one day and the same assessees as Assessor C on another day.

TABLE 1 Means and standard deviations of dimension and exercise ratings received by (a) White and non-White and for (b) male and female assessees in three exercises

	Tactical analysis		In-box		Oral presentation		Average	
	White	Non-White	White	Non-White	White	Non-White	White	Non-White
	N = 73	*N* = 124	*N* = 73	*N* = 124	*N* = 73	*N* = 124	*N* = 73	*N* = 124
(a) Overall dimension × Exercise ratings by assessee race								
CR	3.44 (0.84)	3.32 (0.96)	3.36 (1.10)	3.08 (1.10)	3.57 (0.87)	3.42 (0.92)	3.46 (0.61)	3.27 (0.67)
CS	3.28 (0.94)	3.22 (0.98)	3.46 (1.11)	3.30 (1.05)	3.83 (0.89)	3.67 (0.88)	3.52 (0.61)	3.39 (0.67)
LD	3.59 (0.91)	3.40 (0.95)	3.55 (1.10)	3.36 (1.16)	3.72 (0.91)	3.46 (0.94)	3.62 (0.63)	3.41 (0.67)
OC	3.68 (0.82)	3.45 (0.83)	3.73 (0.92)	3.56 (0.89)	3.80 (0.89)	3.57 (0.87)	3.74 (0.59)	3.53 (0.65)
PS	3.53 (0.89)	3.46 (0.90)	3.38 (1.09)	3.07 (1.09)	3.75 (0.82)	3.54 (0.92)	3.55 (0.61)	3.36 (0.67)
WC	3.33 (0.91)	3.45 (0.95)	3.18 (1.02)	3.04 (0.90)	3.64 (0.99)	3.48 (0.88)	3.38 (0.74)	3.32 (0.66)
Average	3.47 (0.73)	3.38 (0.79)	3.44 (0.93)	3.23 (0.93)	3.72 (0.79)	3.52 (0.80)	3.55 (0.55)	3.38 (0.61)
	Tactical analysis		In-box		Oral presentation		Average	
	Male	Female	Male	Female	Male	Female	Male	Female
	N = 156	*N* = 33	*N* = 156	*N* = 33	*N* = 156	*N* = 33	*N* = 156	*N* = 33
(b) Overall dimension × Exercise ratings by assessee gender								
CR	3.32 (0.89)	3.48 (0.88)	3.16 (1.09)	3.32 (1.03)	3.43 (0.93)	3.63 (0.79)	3.30 (0.66)	3.48 (0.60)
CS	3.21 (0.95)	3.39 (0.92)	3.32 (1.07)	3.50 (0.95)	3.65 (0.91)	4.05 (0.69)	3.40 (0.66)	3.65 (0.60)
LD	3.43 (0.91)	3.63 (0.93)	3.42 (1.14)	3.52 (1.04)	3.50 (0.97)	3.73 (0.78)	3.45 (0.67)	3.63 (0.62)
OC	3.52 (0.82)	3.59 (0.76)	3.57 (0.88)	3.89 (0.87)	3.60 (0.91)	3.88 (0.73)	3.56 (0.63)	3.79 (0.61)
PS	3.46 (0.87)	3.52 (0.92)	3.19 (1.14)	3.15 (0.98)	3.53 (0.91)	3.89 (0.66)	3.40 (0.67)	3.52 (0.58)
WC	3.30 (0.91)	3.81 (0.84)	3.01 (0.93)	3.39 (0.98)	3.45 (0.96)	3.88 (0.67)	3.25 (0.69)	3.69 (0.59)
Average	3.28 (0.95)	3.46 (0.85)	3.37 (0.77)	3.57 (0.76)	3.53 (0.84)	3.84 (0.60)	3.39 (0.60)	3.63 (0.53)

Notes. CR = Conflict Resolution/Teamwork; CS = Customer Service; LD = Leadership; OC = Oral Communication; PS = Problem Solving/Decision Making; WC = Written Communication.

TABLE 2 Means and standard deviations of dimension and exercise ratings given by (a) White and non-White and (b) male and female assessors in three exercises

	Tactical analysis		In-box		Oral presentation		Average	
	White	Non-White	White	Non-White	White	Non-White	White	Non-White
	N = 19	*N* = 27	*N* = 20	*N* = 27	*N* = 20	*N* = 27	*N* = 20	*N* = 27
(a) Overall dimension × Exercise ratings by assessor race								
CR	3.35 (0.35)	3.38 (0.45)	3.10 (0.46)	3.27 (0.54)	3.32 (0.35)	3.65 (0.50)	3.25 (0.26)	3.43 (0.38)
CS	3.25 (0.46)	3.22 (0.45)	3.32 (0.54)	3.43 (0.48)	3.54 (0.33)	3.89 (0.48)	3.38 (0.26)	3.51 (0.39)
LD	3.49 (0.33)	3.48 (0.48)	3.32 (0.55)	3.56 (0.60)	3.41 (0.37)	3.72 (0.50)	3.40 (0.28)	3.59 (0.43)
OC	3.44 (0.32)	3.57 (0.34)	3.54 (0.43)	3.69 (0.38)	3.49 (0.36)	3.80 (0.46)	3.49 (0.24)	3.69 (0.34)
PS	3.50 (0.36)	3.48 (0.41)	3.08 (0.47)	3.29 (0.56)	3.48 (0.36)	3.76 (0.50)	3.34 (0.26)	3.51 (0.39)
WC	3.30 (0.31)	3.38 (0.51)	3.11 (0.41)	3.09 (0.39)	3.45 (0.51)	3.66 (0.42)	3.27 (0.27)	3.37 (0.34)
Average	3.34 (0.30)	3.42 (0.39)	3.24 (0.43)	3.39 (0.45)	3.45 (0.33)	3.75 (0.45)	3.36 (0.23)	3.52 (0.36)
	Tactical analysis		In-box		Oral presentation		Average	
	Male	Female	Male	Female	Male	Female	Male	Female
	N = 38	*N* = 11	*N* = 38	*N* = 11	*N* = 38	*N* = 11	*N* = 38	*N* = 11
(b) Overall dimension × Exercise ratings by assessor gender								
CR	3.41 (0.38)	3.21 (0.45)	3.20 (0.47)	3.20 (0.62)	3.52 (0.45)	3.43 (0.52)	3.38 (0.33)	3.28 (0.40)
CS	3.26 (0.42)	3.12 (0.54)	3.37 (0.47)	3.46 (0.58)	3.72 (0.41)	3.83 (0.55)	3.45 (0.30)	3.48 (0.45)
LD	3.51 (0.39)	3.36 (0.48)	3.46 (0.57)	3.40 (0.64)	3.59 (0.43)	3.57 (0.60)	3.52 (0.34)	3.44 (0.49)
OC	3.53 (0.33)	3.52 (0.35)	3.62 (0.38)	3.66 (0.48)	3.67 (0.40)	3.70 (0.58)	3.60 (0.28)	3.62 (0.42)
PS	3.51 (0.35)	3.40 (0.48)	3.22 (0.50)	3.13 (0.63)	3.65 (0.42)	3.62 (0.58)	3.46 (0.31)	3.37 (0.45)
WC	3.40 (0.38)	3.19 (0.59)	3.12 (0.38)	3.05 (0.44)	3.59 (0.44)	3.50 (0.54)	3.37 (0.26)	3.23 (0.43)
Average	3.44 (0.32)	3.30 (0.42)	3.33 (0.41)	3.32 (0.52)	3.62 (0.39)	3.61 (0.53)	3.46 (0.28)	3.40 (0.42)

Notes. CR = Conflict Resolution/Teamwork; CS = Customer Service; LD = Leadership; OC = Oral Communication; PS = Problem Solving/Decision Making; WC = Written Communication.

To model the structure of these complex data accurately, we used a cross-classified multilevel modeling approach (Fielding & Goldstein, 2006; Hox, 2002). Multilevel models are useful for analyzing nested data, as they allow dependencies among observations to be modeled explicitly and they are flexible for handing groups with varying numbers of observations. A cross-classified model, in particular, allows observations to be nested in more than one way. A common example (cf. Fielding & Goldstein, 2006) is of students grouped within neighborhoods and also within schools; each student belongs to one neighborhood and one school, but students in the same neighborhood may attend different schools and students attending the same school may live in different neighborhoods. In the present study, we were able to model dimension ratings as being nested within assessees and also within assessors, even though assessees were not perfectly nested within assessors and assessors were not perfectly nested within assessees. Figure 1 provides an illustration of the model. The dependent variable of interest in this model is the individual rating made by one assessor for one assessee on one dimension. These ratings are grouped within assessees (as each assessee was rated on multiple dimensions), but they are also grouped within assessors (as each assessee was rated by two assessors). A full explanation of the cross-classified multilevel model estimation is contained in the Appendix.

We modeled race and gender effects separately for each exercise, which resulted in a total of six sets of cross-classified models. Each set of models were analyzed by introducing model estimates in the following order: fixed intercepts only (Model 1), assessee and assessor race (or gender) fixed main effects (Model 2), assessee–assessor race (gender) fixed interaction effect (Model 3), and assessor race (gender) random effects (Model 4). Cross-classified analyses were conducted using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the *R* statistical software (Version 1.0.143; R Core Development Team, 2016). Following the recommendations of Bates (2010) and Hox (2002), we used a model comparison approach to evaluate the "significance" of specific effects. Although the lme4 package provides standard errors for individual model parameters, Bates (2006) notes that using these to evaluate the statistical significance of these parameters is problematic, because of concerns about the appropriate degrees of freedom for these tests. Instead, both Bates (2010) and Hox (2002) recommend comparing the overall fit of nested models, using the deviance statistic, to determine whether including a given parameter substantially improves the fit of the model. Thus, we compared the deviance for each model against the deviance value for the previous model. If the more complex model fit significantly better than the less complex model, we considered the

TABLE 3 Ratings given by White and non-White assessors to White and non-White assessees

		Assessee's race					
		Tactical analysis		In-box		Oral presentation	
	Assessor race	White	Non-White	White	Non-White	White	Non-White
Number of ratings	White	57	93	68	97	53	117
	Non-White	89	155	78	151	93	131
Dimension							
CR	White	3.39	3.24	3.26	2.97	3.49	3.28
	Non-White	3.47	3.36	3.44	3.15	3.62	3.55
CS	White	3.19	3.23	3.43	3.20	3.74	3.48
	Non-White	3.34	3.21	3.48	3.36	3.88	3.84
LD	White	3.50	3.39	3.51	3.15	3.67	3.31
	Non-White	3.64	3.41	3.58	3.49	3.75	3.60
OC	White	3.63	3.31	3.68	3.45	3.79	3.39
	Non-White	3.72	3.53	3.78	3.63	3.81	3.74
PS	White	3.52	3.40	3.35	2.88	3.64	3.41
	Non-White	3.54	3.49	3.40	3.20	3.82	3.65
WC	White	3.36	3.28	3.27	2.96	3.56	3.36
	Non-White	3.30	3.55	3.10	3.09	3.69	3.58
Average	White	3.43	3.31	3.42	3.10	3.65	3.37
	Non-White	3.50	3.43	3.46	3.32	3.76	3.66

Notes. CR = Conflict Resolution, CS = Customer Service, LD = Leadership, OC = Oral Communication, PS = Problem Solving, WC = Written Communication.

TABLE 4 Ratings given by male and female assessors to male and female assessees

		Assessee's gender					
		Tactical analysis		In-box		Oral presentation	
	Assessor gender	Male	Female	Male	Female	Male	Female
Number of ratings	Male	248	64	245	67	236	76
	Female	53	13	50	16	48	18
Dimension							
CR	Male	3.35	3.18	3.18	3.09	3.45	3.38
	Female	3.58	3.12	3.3	3.38	3.71	3.42
CS	Male	3.21	3.22	3.32	3.34	3.63	3.71
	Female	3.55	2.77	3.5	3.5	4.05	4.06
LD	Male	3.46	3.3	3.45	3.29	3.49	3.53
	Female	3.74	3.19	3.51	3.53	3.8	3.56
OC	Male	3.51	3.58	3.58	3.54	3.6	3.6
	Female	3.63	3.42	3.9	3.84	3.88	3.89
PS	Male	3.48	3.39	3.22	3.1	3.54	3.52
	Female	3.63	3.08	3.17	3.09	3.92	3.83
WC	Male	3.33	3.16	3.03	2.93	3.47	3.36
	Female	3.86	3.62	3.42	3.31	3.88	3.89
Average	Male	3.39	3.3	3.3	3.21	3.53	3.52
	Female	3.66	3.2	3.47	3.44	3.87	3.77

Notes. CR = Conflict Resolution, CS = Customer Service, LD = Leadership, OC = Oral Communication, PS = Problem Solving, WC = Written Communication.

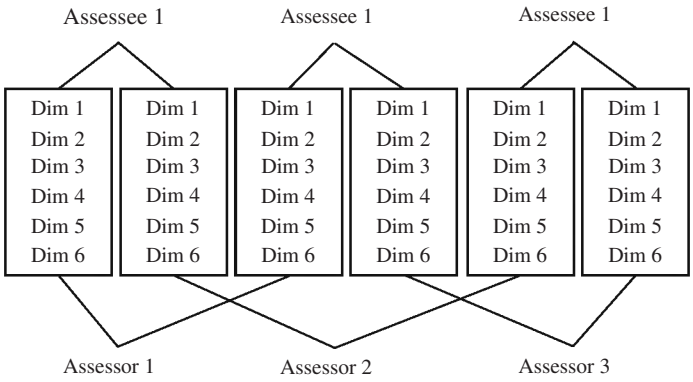

FIGURE 1 Illustration of assessors rating different assessees and assessees being rated by different assessors

parameters added in the more complex model to make a substantive contribution to explaining the overall variance in ratings.

4.5 | Aggregated ratings: Subgroup difference effect sizes and adverse impact analyses

Subgroup differences in the ratings assessees received (assessors gave) due to assessee/assessor race and gender are reported using Cohen's *d* effect size estimates for each exercise and across exercises. Subgroup differences due to assessee race (White/non-White) and gender (male/female) were calculated from PEDRs that were arithmetically derived by averaging the pair of assessors' ratings of each assessee on each dimension within each exercise. From there, the six PEDRs within each exercise were averaged to calculate each assessee's average exercise score and each average exercise score was averaged to calculate each assessee's OAR. This approach reflects the actual scoring procedures used in this and other ACs using statistical integration. Subgroup differences due to *assessor* race and gender were calculated by aggregating each assessor's ratings across the assessees the assessor rated within each exercise. This resulted in separate within-assessor ratings, aggregated across all observed assessees, for each dimension within each exercise. From there, the aggregation process mirrored that used to compute subgroup difference effect sizes for assessee race and gender. All assessors rated at least four assessees in each of the three exercises, with the majority of assessors rating 10 assessees within each exercise. The mean number of assessees rated across all three exercises were: $White_M$ = 24.17; non-$White_M$ = 24.00; $Male_M$ = 24.16; $Female_M$ = 23.87). Final promotion decisions were examined for the presence of adverse impact against White or non-White and, separately, male or female assessees using the 4/5th (80%) rule and chi-square test (Bobko & Roth, 2010).

5 | RESULTS

5.1 | Cross-classified multilevel models for individual-level leniency and similar-to-me bias

Table 5 presents a comparison of the fit statistics for each of the six sets of models analyzed to test for leniency and similar-to-me bias. Table 6 summarizes the estimates of all possible effects (Model 4) for each of the six analyses. Individual parameter estimates changed very little (<0.01) from one model to the next and reporting the results of the most complex model was done to balance completeness with brevity.[1] Interclass correlations (ICCs) estimated from Table 6 indicate that, across the six sets of models, the variance in dimension ratings accounted for by assessee effects ranged from 67%[2] to 74%, while the variance accounted for by assessor effects was consistently ≤1%. In other words, most of the variance in dimension ratings was due to differences among the assessees, and the assessors were essentially interchangeable.

For *race*, the best-fitting model for the Oral Presentation exercise included the addition of assessor and assessee race main effects (Model 2). The direction of the estimate presented in Table 6 shows that, on average, White assessees received more lenient ratings than non-White assessees. Assessor race demonstrated a trivial effect. The best-fitting model for the Tactical Analysis exercise further included the assessee × assessor race fixed interaction effect (Model 3), suggesting similar-to-me bias among non-White assessees rated by non-White assessors. The best-fitting model for the In-box exercise included all assessee and assessor fixed and random race effects (Model 4). The random effects for *assessee race* did improve model fit, suggesting that there is detectable variability, but the size of this variance was again quite small (<2%). Given the trivial variance attributable to assessors, as estimated by ICCs, this is not surprising. Assessors' idiosyncratic biases appeared to play little role in their ratings.

For *gender*, the results were more straightforward. The best models were those including only the intercepts (Model 1). These indicate that, consistent with much other AC research (Lievens & Conway, 2001), rating dimensions varied in difficulty within and across exercises. Adding main effects, interactions, or random effects by either assessee or assessor gender did not significantly improve the fit of these models. These findings suggest that assessee and assessor gender (and their interaction) played a minimal role in the commission of leniency or similar-to-me bias.

Taken together, cross-classified multilevel modeling results provided limited support for hypotheses regarding the systematic individual-level effects for leniency error in either assessee or assessor ratings due to race or gender. Systematically greater leniency was

TABLE 5 Comparisons of model fit for cross-classified multilevel models testing leniency and similar-to-me error

	Tactical exercise						In-box						Oral presentation					
	Dev	AIC	*df*	Δ Dev	Δ *df*	*p*	Dev	AIC	*df*	Δ Dev	Δ *df*	*p*	Dev	AIC	*df*	Δ Dev	Δ *df*	*p*
Models for race																		
Model 1: Intercepts only	4,277.0	4,295.0	9	–	–	–	4,400.5	4,418.5	9	–	–	–	4,108.2	4,126.2	9	–	–	–
Model 2: Fixed main effects (assessor race, assessee race)	4,273.4	4,295.4	11	3.6	2	0.17	4,396.0	4,418.0	11	4.5	2	0.11	**4,102.4**	**4,124.4**	**11**	**5.8**	**2**	**0.06**
Model 3: Fixed interaction (assessor race × assessee race)	**4,268.7**	**4,292.7**	**12**	**4.7**	**1**	**0.03**	4,394.5	4,418.5	12	1.5	1	0.22	4,101.2	4,125.2	12	1.2	1	0.27
Model 4: Random effects by assessor	4,267.8	4,301.8	17	0.9	5	0.97	**4,381.1**	**4,415.1**	**17**	**13.4**	**5**	**0.02**	4,100.5	4,134.5	17	0.7	5	0.98
Models for gender																		
Model 1: Intercepts only	**4,111.7**	**4,129.7**	**9**	–	–	–	**4,242.3**	**4,260.3**	**9**	–	–	–	**3,953.6**	**3,971.6**	**9**	–	–	–
Model 2: Fixed main effects (assessor gender, assessee gender)	4,109.3	4,131.3	11	2.4	2	0.30	4,241.2	4,263.2	11	1.1	2	0.58	3,949.0	3,971.0	11	4.6	2	0.10
Model 3: Fixed interaction (assessor gender × assessee gender)	4,108.8	4,132.8	12	0.5	1	0.48	4,240.7	4,264.7	12	0.5	1	0.48	3,949.0	3,973.0	12	0.0	1	1.0
Model 4: Random effects by assessor	4,108.2	4,142.2	17	0.6	5	0.99							3,946.8	3,980.8	17	2.2	5	0.82

Notes. Dev = deviance. *df* = degrees of freedom. Values in **bold text** indicate the simplest acceptable model for each exercise, beyond which adding further parameters did not improve fit.

TABLE 6 Parameter estimates for Model 4 of cross-classified multilevel models testing leniency and similar-to-me error across exercises

	Race						Gender					
	Tactical Analysis		In-box		Oral Presentation		Tactical Analysis		In-box		Oral Presentation	
Fixed effects	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Intercepts (for each dimension)												
CR	3.44	0.10	3.31	0.11	3.57	0.11	3.32	0.10	3.15	0.10	3.42	0.07
CS	3.31	0.10	3.48	0.11	3.83	0.11	3.22	0.10	3.32	0.10	3.68	0.07
LD	3.54	0.10	3.56	0.11	3.66	0.11	3.43	0.10	3.40	0.10	3.50	0.07
OC	3.61	0.10	3.75	0.11	3.76	0.11	3.51	0.10	3.59	0.10	3.60	0.07
PS	3.56	0.10	3.31	0.11	3.71	0.11	3.44	0.10	3.15	0.10	3.55	0.07
WC	3.48	0.10	3.22	0.11	3.64	0.11	3.36	0.10	3.04	0.10	3.48	0.07
Predictors												
Assessee race (gender)	−0.18	0.12	−0.25	0.14	**−0.22**	**0.12**	0.19	0.15	0.20	0.18	0.31	0.15
Assessor race (gender)	−0.02	0.06	<0.01	0.05	0.04	0.06	−0.05	0.05	0.02	0.05	−0.03	0.06
Assessee race (gender) × Assessor race (gender)	**0.15**	**0.07**	0.08	0.06	0.06	0.06	0.08	0.10	−0.06	0.09	−0.01	0.09
Random effects												
σ^2 in intercepts by assessee	0.56		0.83		0.60		0.56		0.84		0.60	
σ^2 in intercepts by assessor	<0.01		<0.01		0.01		0.01		<0.01		0.01	
Residual σ^2 in intercepts	0.27		0.28		0.25		0.27		0.28		0.25	
σ^2 in effects of assessee race (gender) by assessor	<0.01		**0.02**		<0.01		<0.01		[a]		0.01	
σ^2 in effects of assessor race by assessor	<0.01		**0.01**		0.04		0.02		[a]		0.04	

Notes. CR = Conflict Resolution, CS = Customer Service, LD = Leadership, OC = Oral Communication, PS = Problem Solving, WC = Written Communication. SE = standard error. Values in **bold type** indicate parameters that improved the fit of the overall model (see Table 5).

[a]Model including random effects by gender did not converge for the In-box; coefficients above are from Model 3.

found to be associated with ratings received by White assessees, relative to non-White assessees. This supported H1a. No such differences were found in ratings received by female and male assessees, which failed to support H1b. Similarly, no differences in leniency were found in the ratings given by assessors, based on either assessor race or gender. Thus, H1c and H1d were also not supported. Mixed support was found for the systematic individual-level effects for similar-to-me bias. Cross-classified results showed same-race effects among pairs of non-White assessees and assessors, which partially supported H2a. No similar-to-me effects were found with regard to gender, which failed to support H2b.

5.2 | Effect size estimates for mean subgroup differences and adverse impact

Findings for assessee and assessor race and gender subgroup difference effect sizes are presented in Table 7. We begin by discussing assessee subgroup differences. For perspective, these effect sizes (*d*s) are relatively small (i.e., less than 0.50, or half a standard deviation). Consistent with the cross-classified results of individual-level data, subgroup-level mean differences showed White assessees were rated higher than their non-White counterparts across exercises (H1a). In each exercise, differences were smaller in magnitude than either the Black–White or Hispanic–White mean differences observed in Dean et al.'s (2008) meta-analysis. Subgroup-level mean differences also showed female assessees were consistently rated higher than male assessees across all three exercises. While these findings are inconsistent with the null individual-level effects observed through the cross-classified analyses, they are consistent with the mean differences proposed in H1b. Assessee gender mean differences observed in the Tactical Analysis and In-box exercises were similar those estimated in Dean et al.'s meta-analysis, while the mean difference observed for the Oral Presentation was larger (d = 0.40).

Findings regarding assessor race subgroup-level mean differences are inconsistent with the cross-classified null results of individual-level data: they showed non-White assessors gave higher ratings than their White counterparts across exercises, with effect sizes ranging from small (Tactical d = −0.23 and In-box d = −0.35) to large (Oral Presentation d = −0.76). This finding is, however, consistent with H1c. Subgroup-level differences showing that male assessors gave higher ratings than female assessors are inconsistent with the null results of individual-level cross-classified analyses, as well as inconsistent with H1d, which proposed that female assessors would provide higher ratings.

These generally small mean subgroup differences did not result in adverse impact in promotion decisions. Using the 4/5th (80%) rule, the percent of the racial or gender subgroup with the smaller promotion rate was not less than 80% percent of the promotion rate for the subgroup with the larger promotion rate. Using the chi-square test for the distribution of pass and fail found no statistically significant differences for either gender or race.

TABLE 7 Effect sizes (*d*) comparing ratings given to White versus Non-White and male versus female assessees by White versus Non-White and male versus female assessors

	Tactical	In-box	Oral presentation
Assessee			
Race	0.12	0.23	0.25
Gender	−0.20	−0.26	−0.40
Assessor			
Race	−0.23	−0.35	−0.76
Gender	0.42	0.23	0.02

Note. Negative (-) *d*s indicate higher ratings for female assessees and non-White assessors.

5.3 | Dimensions and exercises

Results related to Research Question A show some differences in the results across exercises for race effects in leniency and similar-to-me bias. However, with one exception, the effect sizes were small. For the Tactical Analysis, 67% of the variance in rating was due to differences in assessee performance, whereas less than 1% of the variance in ratings was due to assessors. For the interaction, Non-White assessees rated by non-White assessors received slightly higher ratings. This interaction effect, while relatively small (15%) in comparison with overall assessee effects, warrants attention and further study of technical exercises. For the In-box, 75% of the variance in ratings was due to assessees, whereas less than 1% due to assessors and trivial amounts (1% and 2%) due to interactions of assessors and assessees. For the Oral Presentation exercise, 70% of the variance was due to assessees. While Whites received higher ratings than non-Whites, the variance due to assessors was less than 1%. No differences in results across exercises were found for gender effects.

Related to Research Question B, examination of results in Table 6 suggests that dimensions differ in terms of difficulty within exercises and across exercises. Constraints on the analyses precluded determination of significance of these differences, their magnitude, or interactions with demographics of assessee or assessor. More research is warranted to study rating errors across dimensions and exercises.

6 | DISCUSSION

In the present study, we used individual-level assessor ratings to examine the differential presence of leniency and similar-to-me bias in assessee and assessor race and gender subgroups. Our findings support only two of the hypothesized individual-level effects: (a) greater leniency toward White assessees, and (b) similar-to-me bias among non-White assessor–assessee pairs. No other individual-level effects were statistically significant. Positing that even small individual-level effects can "trickle up" to produce more pronounced subgroup-level effects, we subsequently examined mean subgroup

differences and assessed for the presence of adverse impact. These analyses showed effects in the direction consistent with hypothesized assessee race and gender and assessor race effects. No evidence of adverse impact, due to race or gender, was found.

Perhaps the strongest support for the "trickle-up" explanation comes from findings regarding assessee race effects. Individual-level ratings in the Oral Presentation exercise showed a statistically significant effect for assessee race, suggesting that White assessees systematically received more lenient ratings than non-White assessees. This effect translated into a larger subgroup difference estimate between White and non-White assessees in that exercise. While individual-level effects for assessee gender were not statistically significant in any of the three exercises, the magnitude of individual-level and subgroup-level assessee gender effects across exercises was generally similar to those observed for assessee race effects. Moreover, exercises showing larger individual-level effects of assessee race and gender also showed larger subgroup-level mean differences. Individual-level and subgroup-level effects were consistently in the direction we hypothesized, consistent with prior theory and research, even when these effects were small and did not reach the threshold for significance.

However, no such pattern of "trickle-up" effects was evident for effects based on assessor race or gender. Individual-level effects for both assessor race and gender were even smaller than individual-level assessee effects and essentially zero, yet assessor subgroup differences were as large as (or larger than) assessee subgroup differences. Moreover, the direction of individual-level assessor effects was not entirely consistent with the direction of subgroup-level assessor effects. Thus, although there was evidence that different assessors engaged in rating errors to different extents, these errors were not associated with the assessors' race or gender and did not aggregate in any meaningful way to explain differences between assessor subgroups.

Despite mixed support for the proposed "trickle-up" effect from individual-level to subgroup-level differences, across exercises the magnitude of subgroup differences, themselves, was generally small, with Cohen's *d* estimates similar to or smaller than those observed in Dean et al.'s (2008) meta-analysis. Conclusions based on the present study are more consistent with early conclusions that subgroup differences in ACs are smaller than paper-and-pencil tests (Schmitt & Mills, 2001), and they are in contrast with more recent reviews highlighting the upper bounds of assessees subgroup differences found in some studies of ACs. Perhaps most importantly, neither assessee gender nor race subgroup differences in promotion rates showed evidence of adverse impact. These findings are particularly relevant given that these data came from an AC used for high-stakes promotional testing in an occupation that has traditionally been dominated by White males.

What might explain these findings (or lack thereof)? The simplest explanation for these findings is that assessors effectively rated assessees on job-related behaviors, as opposed to demographic characteristics. In this promotional assessment program, assessees are pre-screened for satisfactory job performance and basic knowledge of the job; they engaged in highly job-relevant assessment tasks; and assessors are trained to observe multiple behaviors related to multiple, job-relevant dimensions of performance effectiveness. While the job and location, namely police work in a major metropolitan city, might be a setting where biases related to gender and race would operate, they had trivial overall effects for this sample of assessees and assessors. Thus, based on the results of this study, coupled with the extensive evidence summarized elsewhere (e.g., Thornton et al., 2015), variations in these assessment center ratings would appear to be a function of behavioral evidence of performance, not the demographic variables of race or gender. It is quite possible that assessee–assessor similarity can bias subjective evaluations, but surface-level characteristics such as race and gender may be less of a factor than deep-level characteristics such as personality and value systems (Bell, Villado, Lukasik, Belau, & Briggs, 2011). However, in this assessment center, the assessors were not familiar with the assessees, and thus deep-level characteristics, which often require some level of familiarity to perceive, were unlikely to operate.

This explanation has important implications for broader issues raised regarding variability in and magnitude of subgroup differences observed across AC studies (Bobko & Roth, 2013; Dean et al., 2008; Ployhart & Holtz, 2008). The *Guidelines and Ethical Considerations for Assessment Center Operations*, first published in 1975 and most recently revised in 2015 (International Task Force, 2015) outlines best practices for designing, implementing, and managing various components inherent to the AC method, including: identifying job-relevant behavioral constructs; developing simulation exercises that sufficiently elicit job-relevant behaviors; and training AC assessors. When ACs are developed in close accordance with these *Guidelines*, findings of subgroup differences *should* be consistently small in magnitude. Conversely, a range of issues can arise throughout the design and execution phases that can enhance threats to AC validity due to rater error. Examples include failing to adequately train assessors, asking assessors to work long hours; not giving assessors sufficient time to evaluate and score assessees; and not effectively using post-exercise assessor discussions to eliminate individual rating biases and errors from post-consensus ratings (Caldwell et al., 2003; Dewberry, 2011; Dewberry & Jackson, 2016).

The population of ACs reviewed by Bobko and Roth (2013) and meta-analyzed by Dean et al. (2008) included ACs that likely differed in rigor and quality of implementation. The AC from which the data in the current study came adhered closely to as many of the *Guidelines*' best practices as was practically (or legally) feasible in this high-stakes promotion setting. Thus, the lack of effects observed at the individual and subgroup levels and the absence of adverse impact may come as little surprise. Because assessment centers are a method, with rating dimensions and simulations tailored to the job requirements of specific jobs, AC designs are far more variable as compared to off-the-shelf paper-and-pencil tests. This variability introduces greater opportunity for issues involving evaluative consistency, accuracy, and fairness to arise.

The present findings do not discount the concerns raised in recent reviews over the actual or perceived fairness associated with

ACs. Rather, recent criticisms of the AC method may be indicative of the growth in popularity ACs have experienced over the past 40 years, as a result of early evidence establishing relatively strong criterion-related validity along with minimal potential for adverse impact. Unfortunately, with this increased popularity comes increased potential for less rigorously designed ACs to be implemented, which can have severe consequences for the results of a specific AC (both in terms of criterion-related validity and adverse impact), as well as raise questions over the fairness of the method as a whole. The present findings: (a) reaffirm that, when developed and implemented in close accordance to established best practices, ACs can minimize bias and the potential for adverse impact and (b) remind AC users that great care must be taken when developing and implementing ACs.

6.1 | Limitations and future research directions

Although this study presents some of the first evidence about assessor race and gender main effects and assessee–assessor interaction effects, and the first evidence examining how individual-level rating errors translate into subgroup-level differences, the nature of these field data from a high-stakes promotional testing setting also present some limitations. First, the setting did not allow for equal numbers of male/female or White/non-White assessors or assessees, nor a fully crossed design in which all assessors rated all assessees. Such features would have enhanced our statistical power and ability to rule out alternative explanations.

Similarly, racial subgroup comparisons were limited to White versus non-White (i.e., Black *and* Hispanic) categories, which minimized precision for identifying more specific same-race biases. Even after combining Black and Hispanic subgroups, the sample sizes for some comparisons remained small, which limited statistical power for identifying interactions between assessor and assessee subgroups (Murphy & Russell, 2017), and we were unable to examine intersections of subgroups–whether, for example, Black female (or White male) assessors showed bias in favor of Black female (White male) assessees. It is possible that similar-to-me bias becomes more salient as the number of qualities on which assessors and assessees are similar increases. Given the limited evidence that exists regarding either same-race or same-gender effects, future research examining assessee–assessor similarity on multiple demographic or psychological characteristics is warranted.

Second, although our study offers insight into the role of individual-level assessor errors in explaining subgroup differences in AC outcomes, we did not directly test the explanations we proposed for the errors themselves (e.g., stereotypes, similarity-attraction). In an operational AC, questioning assessors about their endorsement of stereotypes or their perceptions of similarity to individual candidates would likely be viewed as burdensome at best; at worst, it might cause reactivity or lead to perceptions of greater unfairness. However, other techniques, such as content analysis of assessors' narrative comments (which were not used in the present AC) or debriefing interviews with assessors after the AC could shed light on these explanatory mechanisms in future research.

Third, an alternative explanation for the small subgroup differences we found could be that the dimensions and exercises in this operational AC did not vary enough, or were not sufficiently cognitively loaded to allow differences to emerge. In other settings, evidence has demonstrated that assessee subgroup differences vary as a function of the cognitive and job knowledge components of exercises (Roth et al., 2008; see also Ployhart & Holtz, 2008). Therefore, future research is needed to more clearly understand which, to what extent, and how exercise characteristics may systematically impact rating errors, and assessor and assessee subgroup differences.

7 | CONCLUSION

Recent research has called into question the assertion that the AC method typically produces small subgroup differences and little to no adverse impact. Differences in claims highlight the need for research aimed at understanding the mechanisms through which subgroup differences arise. Using data from a high-stakes promotional AC, the present study used cross-classified multilevel modeling to assess previously unstudied individual-level assessor biases. Findings provide some evidence of the "trickle up" effects of individual-level biases on subgroup-level mean differences. Findings also contribute much needed evidence regarding assessor subgroup differences and assessee–assessor similarity effects. Findings also highlight the importance of rigorous AC development and implementation as a means of minimizing subgroup differences and the potential for adverse impact. Finally, we propose that our largely null individual-level effects and small subgroup-level differences are due to the most straightforward possibility: assessor ratings from this AC, which closely followed best-practice design and implementation guidelines, effectively differentiated on the basis of job-related behaviors as opposed to assessee demographic characteristics.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions to this study. As this paper involves the work of four generations of assessment center scholars, order of authorship was based on who begat whom in academic advising and mentoring.

ENDNOTES

[1]A full set of model results is available from the authors upon request.

[2]For Tactical, the sum of variance attributable to assessors (0.56) + assessors (<0.01) + residual (0.27) = 0.83; 0.56/0.83 = 0.67.

REFERENCES

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

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. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Barrett, G. V., Doverspike, D., & Young, C. M. (2010). The special case of public sector police and fire selection. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 437–462). San Francisco, CA: Jossey-Bass.

Bates, D. (2006). R] lmer, p-values and all that. Retrieved from https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Retrieved from http://lme4.r-forge.r-project.org/book

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bell, S. T., Villado, A. J., Lukasik, M. A., Belau, L., & Briggs, A. L. (2011). Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *Journal of Management*, *37*, 709–743. https://doi.org/10.1177/0149206310365001

Blair, C. A., Hoffman, B. J., & Ladd, R. T. (2016). Assessment centers vs situational judgment tests: Longitudinal predictors of success. *Leadership and Organizational Development Journal*, *37*, 899–911. https://doi.org/10.1108/LODJ-12-2014-0235

Bobko, P., & Roth, P. L. (2010). An analysis of two methods for assessing and indexing adverse impact: A disconnect between the academic literature and some practice. In J. L. Outtz (Ed.), *Adverse impact* (pp. 29–49). New York, NY: Routledge.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, *66*, 91–126. https://doi.org/10.1111/peps.12007

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in judgment of human performance. *Organizational Behavior and Human Performance*, *20*, 238–252.

Bowen, C.-C., Swim, J., & 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.

Buckett, A., Becker, J. R., & Roodt, G. (2017). General performance factors and group differences in assessment center ratings. *Journal of Managerial Psychology*, *32*, 298–313. https://doi.org/10.1108/JMP-08-2016-0264

Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.

Caldwell, C., Thornton, G. C. III, & Gruys, M. L. (2003). Ten classic assessment center errors: Challenges to selection validity. *Public Personnel Management*, *32*, 73–88. https://doi.org/10.1177/009102600303200104

Coleman, J. L. (1992). *Police assessment testing: An assessment center handbook for law enforcement personnel* (2nd ed.). Springfield, IL: Charles C Thomas.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218–244. https://doi.org/10.1037/0033-2909.90.2.218

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685–691. https://doi.org/10.1037/0021-9010.93.3.685

Dewberry, C. (2011). Integrating candidate data: Consensus or Arithmetic. In N. Povah & G. C. Thornton (Eds.), *Assessment and development centres: Strategies for global talent management* (pp. 97–113). Farnham, England: Gower.

Dewberry, C., & Jackson, D. J. R. (2016). The perceived nature and incidence of dysfunctional assessment center features and processes. *International Journal of Selection and Assessment*, *24*, 189–196. https://doi.org/10.1111/ijsa.12140

Eagly, A., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Eagly, A., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin*, *15*, 543–558. https://doi.org/10.1177/0146167289154008

Elsesser, K. M., & Lever, J. (2011). Does gender bias against female leaders persist? Quantitative and qualitative evidence from a large-scale survey. *Human Relations*, *64*, 1555–1578.

Falk, A., & Fox, S. (2014). Gender and ethnic composition of assessment centers and its relationship to participants' success. *Journal of Personnel Psychology*, *13*, 11–20. https://doi.org/10.1027/1866-5888/a000100

Fiedler, A. M. (2001). Adverse impact on Hispanic job applicants during assessment center evaluations. *Hispanic Journal of Behavioral Sciences*, *23*, 102–110. https://doi.org/10.1177/0739986301231007

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Research Report No. RR791). London: Department for Education and Skills. Retrieved from https://dera.ioe.ac.uk/6469/1/RR791.pdf

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*, 493–511. https://doi.org/10.1037/0021-9010.72.3.493

Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., … Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, *86*, 713–728. https://doi.org/10.1037/0022-3514.86.5.713

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. https://doi.org/10.1111/j.1744-6570.1998.tb00729.x

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Griffeth, R. W., & Bedeian, A. G. (1989). Employee performance evaluations: Effects of ratee age, rater age, and rater gender. *Journal of Organizational Behavior*, *10*, 83–90.

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior*, *5*, 269–298.

Hoffman, C. C., & Thornton, G. C. III. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*, 455–470. https://doi.org/10.1111/j.1744-6570.1997.tb00916.x

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology*, *29*, 13–30. https://doi.org/10.1111/j.1744-6570.1976.tb00398.x

International Task Force on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment*, *17*, 243–254.

Kane, J. S. (2000). Accuracy and its determinants in distributional assessment. *Human Performance*, *13*, 47–84. https://doi.org/10.1207/S15327043HUP1301_3

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, *28*, 280–290. https://doi.org/10.1037/h0074049

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms.

Psychological Bulletin, *137*, 616–642. https://doi.org/10.1037/a0023557

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*, 56–65. https://doi.org/10.1037/0021-9010.70.1.56

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107. https://doi.org/10.1037/0033-2909.87.1.72

Landy, F. J., & Vasey, J. (1991). Job analysis: The composition of SME samples. *Personnel Psychology*, *44*, 27–50. https://doi.org/10.1111/j.1744-6570.1991.tb00689.x

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. https://doi.org/10.1037/0021-9010.86.6.1202

Lin, T.-R., Dobbins, G. H., & Farh, J.-L. (1992). A field study of race and age similarity effects on interview ratings in conventional interviews. *Journal of Applied Psychology*, *77*, 363–371.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male–female differences: A computer simulation. *American Psychologist*, *51*, 157–158. https://doi.org/10.1037/0003-066X.51.2.157

McFarland, L. A., Ryan, A. M., Sacco, J. M., & Kriska, S. D. (2004). Examination of structured interview ratings across time: The effects of applicant race, rater race, and panel composition. *Journal of Management*, *30*, 435–452. https://doi.org/10.1016/j.jm.2003.09.004

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, *93*, 1042–1052. https://doi.org/10.1037/0021-9010.93.5.1042

Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment-center performance to management progress of women. *Journal of Applied Psychology*, *60*, 527–529. https://doi.org/10.1037/h0076911

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

Murphy, K. R., & Russell, C. J. (2017). Mend it or end it: Redirecting the search for interactions in the organizational sciences. *Organizational Research Methods*, *20*, 549–573.

Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics*, *111*, 915–941. https://doi.org/10.2307/2946676

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Prewett-Livingston, A. J., Feild, H. S., Veres, J. G. III, & Lewis, P. M. (1996). Effects of race on interview ratings in a situational panel interview. *Journal of Applied Psychology*, *81*, 178–186. https://doi.org/10.1037/0021-9010.81.2.178

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and gender effects on performance ratings. *Journal of Applied Psychology*, *74*, 770–780.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roth, P. L., Bobko, P., McFarland, L. A., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black–white differences in overall and exercise scores. *Personnel Psychology*, *61*, 637–662. https://doi.org/10.1111/j.1744-6570.2008.00125.x

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*, 1435–1447. https://doi.org/10.1037/apl0000236

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: Lawrence Erlbaum.

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. https://doi.org/10.1037/0021-9010.86.3.451

Schreurs, B., Druart, C., Proost, K., & DeWitte, K. (2009). Symbolic attributes and organizational attractiveness: The moderating effects of applicant personality. *International Journal of Selection and Assessment*, *17*, 35–46. https://doi.org/10.1111/j.1468-2389.2009.00449.x

Sears, G. J., & Holmvall, C. M. (2010). The joint influence of supervisor and subordinate emotional intelligence on leader–member exchange. *Journal of Business and Psychology*, *25*, 593–605. https://doi.org/10.1007/s10869-009-9152-y

Sherif, M. (1966). *In common predicament: Social psychology of intergroup conflict and cooperation*. Boston, MA: Houghton Mifflin.

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.

Tajfel, H. (1978). *Differentiation between social groups*. London: Academic Press.

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

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

Thornton, G. C. I. I. I., Rupp, D. R., & Hoffman, B. J. (2015). *Assessment center perspectives for talent management strategies*. London, UK: Routledge.

Tinsley, C. H., Howell, T. M., & Amanatullah, E. T. (2015). Who should bring home the bacon? How deterministic views of gender constrain spousal wage preferences. *Organizational Behavior and Human Decision Processes*, *126*, 37–48. https://doi.org/10.1016/j.obhdp.2014.09.003

Varela, O. E., Cater, J. I., & Michel, N. (2011). Similarity attraction in learning contexts: An empirical study. *Human Resource Development Quarterly*, *22*, 49–68. https://doi.org/10.1002/hrdq.20066

Walsh, J. P., Weinberg, R. M., & Fairfield, M. L. (1987). The effects of gender on assessment centre evaluations. *Journal of Occupational and Organizational Psychology*, *60*, 305–309. https://doi.org/10.1111/j.2044-8325.1987.tb00262.x

How to cite this article: Thornton GC III, Rupp DE, Gibbons AM, Vanhove AJ. Same-gender and same-race bias in assessment center ratings: A rating error approach to understanding subgroup differences. *Int J Select Assess*. 2019;27:54–71. https://doi.org/10.1111/ijsa.12229

APPENDIX

In the cross-classified multilevel analyses, we differentiate between *fixed* and *random* effects (following notation from Fielding & Goldsmith, 2006; Hox, 2002). In a traditional single-level regression model, the intercept (β_0) and slope (β_i) parameters are constant across all observations and there is a single residual error term (e_{ij}):

$$Y_{ij} = \beta_0 + \beta_{X_1} + \ldots + \beta_{Xi} + e_{ij}$$

In a multilevel (or *random coefficient)* model, the intercept and slope parameters for the first-level model are modeled as random variables, which can differ across values of the higher-level grouping variable and may be predictable by other variables at that higher level. For example, the intercept β_0 in the previous equation might be a random variable with its own intercept and variability around that intercept, described by:

$$\beta_0 = \gamma_{00} + u_{0j}$$

where γ_{00} is the overall intercept (the intercept of the intercepts) and u_{0j} is the specific error term for group *j*. If there is little variance in the u_{0i}, then β_0 will be close to the overall intercept γ_{00} for all groups. If there is considerable variance in the u_{0i}, however, then the β_0 for different groups may be quite different, and a model that fails to account for group membership will be inaccurate. Variance in intercepts can be interpreted as a main effect of group membership: a specific value of u_{0i} represents the impact of membership in group *i* on the dependent variable. We are typically less interested in estimating this value for a particular group than in estimating the variance in these values across groups.

In the same way, the slopes of individual predictors may also be modeled as random:

$$\beta_1 = \gamma_{10} + v_{1j}$$

where γ_{10} is the intercept of the slopes (essentially, the average slope) and v_{1j} is the specific error term associated with that slope for group *j*. When there is little variance in the v_{1j}, then γ_{10} is a fairly accurate description of the effect of this predictor for all groups. γ_{10} is called a *fixed effect* because it is consistent across all groups. When the v_{1j} vary, however, then the *effects of the predictor* change based on group membership. The variance in v_{1j} is referred to as a *random effect*.

In the context of our current study, the outcome of interest (Y_{ijk}) is a rating made by one assessor for one assessee on each of six dimensions in one exercise (as mentioned previously, analyses were conducted separately for each exercise to pinpoint any effect and to facilitate interpretation). The first level in our model, then, is the rating level. Ratings are then grouped within both assessees (*j*) and assessors (*k*; both Level 2 variables, cross-classified). Because we expect that different dimensions may vary in difficulty within an exercise (e.g., Lievens & Conway, 2001), we estimated a separate intercept for each dimension. Model 1, the baseline model, contains only the intercepts and the variance around the error terms associated with each grouping variable (assessees and assessors):

$$\text{Model 1: } Y_{ijk} = \gamma_{CR} + \gamma_{CS} + \gamma_{LD} + \gamma_{OC} + \gamma_{PS} + \gamma_{WC} + u_{0j} + v_{0k} + e_{ij}$$

As we are not interested in identifying the effect of a specific assessee or assessor, we focus on estimating the variance of u_{0j} and v_{0k}. We then add predictors for main effects. In this example, Model 2 adds fixed effects for the main effects of assessee race and assessor race, and Model 3 adds their interaction:

$$\text{Model 2: } Y_{ijk} = \gamma_{00} + \gamma_{CR} + \gamma_{CS} + \gamma_{LD} + \gamma_{OC} + \gamma_{PS} + \gamma_{WC} + u_{0j} + v_{0k} + \beta_{Cand\ Race} + \beta_{Assr\ Race} + e_{ij}$$

$$\text{Model 3: } Y_{ijk} = \gamma_{00} + \gamma_{CR} + \gamma_{CS} + \gamma_{LD} + \gamma_{OC} + \gamma_{PS} + \gamma_{WC} + u_{0j} + v_{0k} + \beta_{Cand\ Race} + \beta_{Assr\ Race} + \beta_{Cand\ Race\ x\ Assr\ Race} + e_{ij}$$

In Model 4, we add random effects for the assessee and assessor race coefficients; specifically, we allow for the possibility that these effects may vary by assessor, that is, that some assessors may be more influenced than others by race. This type of variability seems both theoretically plausible and of practical interest; if some, but not all, assessors are biased by their own or an assessee's race, it would be of considerable import to understand the antecedents of such biases and how to counteract them. We did not model random effects that varied by assessee, in part because of the limitations of our sample size to estimate such a complex model and in part because the theoretical implications of such effects are ambiguous. If some assessees are more likely than others to have their ratings affected by race, it is not clear why this might be the case or what organizations could do about it. We repeated the analyses above using assessee and assessor gender in place of assessee and assessor race.