

## ABSTRACT

Title of Dissertation: THE DUALITY OF BIAS: PREDICTORS OF RACIALLY MOTIVATED DIFFERENTIAL TEST FUNCTIONING IN INTERVIEW EVALUATIONS

Juliet Aiken, Doctor of Philosophy, 2011

Dissertation Directed By: Dr. Paul Hanges

Department of Psychology

Despite continued interest in and research on discrimination, the complex nature of the process through which it emerges has not been adequately explored. In the current study, I assessed racially-motivated Differential Test Functioning (DTF) and its drivers in an interview context. Specifically, I investigated patterns of DTF-for, DTF-against, and no DTF across three studies. Moreover, I predicted five patterns of responding using in-group belonging (rater race and ethnic identity), prejudice, and motivation to hide prejudice. Results indicate that patterns of responding indicative of DTF-against blacks, DTF-against whites, and no DTF emerged in both student and adult samples. Additionally, in-group belonging and a motivation to hide prejudice appear to predict bias-against, whereas a low in-group belonging may result in no DTF. Implications for research and practice are discussed.

THE DUALITY OF BIAS: PREDICTORS OF RACIALLY MOTIVATED  
DIFFERENTIAL TEST FUNCTIONING IN INTERVIEW EVALUATIONS

by

Juliet Aiken

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2011

Advisory Committee:  
Professor Paul Hanges, Chair  
Professor Cheri Ostroff  
Professor Tom Wallsten  
Professor Mo Wang  
Professor Robert Lissitz, Dean's Representative

© Copyright by

Juliet Aiken

2011

## ACKNOWLEDGEMENTS

I would like to thank my advisor and committee chair, Dr. Paul Hanges, for your support over the past five years. It has been an honor to work with such a high-warmth, high-competence professor, advisor, and mentor. Thank you for all of the time and energy you have devoted to my education and growth. A large part of what I have learned over the course of my time at Maryland can be directly attributed to your efforts and guidance—the rest of what I have learned can probably be attributed to the indirect effects of your influence and mentorship.

Thank you, also, to my committee: Dr. Cheri Ostroff, Dr. Tom Wallsten, Dr. Robert Lissitz, and Dr. Mo Wang. Your feedback and support throughout the dissertation process were invaluable. It was a pleasure to work with such intelligent individuals, and I know that my dissertation is a stronger piece of scholarship thanks to your input.

I would also like to thank all of the individuals who supported the data collection efforts for this dissertation, including those who shared archival data with me, the wonderful firefighters I spoke with online who worked to advertise my study, the individuals who volunteered to participate in my study for no compensation, and my absolutely fantastic team of research assistants. This dissertation would not exist if not for your efforts.

Thank you to my friends in and out of the psychology program. Your willingness to sit patiently through my attempts to discuss my findings was immeasurably helpful in preparing me for my defense. In particular, thank you, Kate and Dan Kuehn, for letting me discuss my results when we should have been wine tasting. Thanks also to the wonderful scholars at DaSAL (Dr. Kevin O’Grady, Dr. Tracy Tomlinson, Brandi Stupica,

Laura Sherman, and Lauren Minacapelli) for letting me talk about this and other studies when we likely should have been discussing clients. Last, I cannot thank Elizabeth Salmon and Rabiah Muhammad enough for patiently listening to me every time I got excited about my analyses and wanted to share them with someone—even though you were often busy with your own commitments.

Finally, thank you to Zak Hutchinson for your support and encouragement throughout my time as a graduate student, and to my family for your constant support throughout my life. I am so lucky to have such pillars of strength. Thank you for believing in me.

## Table of Contents

List of Tables .....	vi
List of Figures .....	viii
The duality of bias: Predictors of racial bias in interview evaluations .....	1
Racial Discrimination in Evaluations.....	3
Social Influences of Discrimination.....	7
Item Response Theory and Differential Test Functioning .....	8
Study 1 Method.....	16
Participants .....	16
Ratings and Frame-of-Reference Training.....	17
Procedure.....	18
Study 1 Results .....	19
Study 1 Discussion.....	22
Study 2 .....	24
Other Individual Differences that Influence Differential Responding.....	27
Study 2 Method.....	30
Participants .....	30
Stimuli .....	31
Design.....	35
Procedure.....	37
Materials.....	38
Prejudice toward blacks and prejudice toward whites.....	38
Motivation to hide prejudice.....	41
Study 2 Results .....	43
Manipulation Check .....	43
Confounds .....	43
Tests of Hypotheses .....	43
Study 2 Discussion.....	51
Study 3 Method.....	54
Participants .....	54
Design.....	56

Procedure.....	56
Measures.....	57
Prejudice toward blacks and prejudice toward whites.....	57
Motivation to hide prejudice.....	62
Ethnic identity.....	63
Study 3 Results .....	64
Manipulation Check .....	64
Confounds .....	64
Tests of Hypotheses .....	65
Supplementary Analyses .....	76
Discussion.....	78
Limitations and Future Directions.....	82
Theoretical Implications.....	87
Practical Implications .....	88
Conclusion .....	90
References.....	182

## List of Tables

Table 1. Black and White Raters' Assessment of Applicants – Archival Sample .....	116
Table 2. Portrayed Candidate Attractiveness, Masculinity, and Age .....	117
Table 3. T-tests for Applicant Voices .....	118
Table 4. Depiction of Experimental Conditions .....	119
Table 5. RCM Main Effects Models of Confounds Related to the Candidates .....	120
Table 6. RCM Interaction Models of Confounds Related to the Candidates .....	121
Table 7. Maximum Likelihood Rotated Factor Solution - White Referent .....	122
Table 8. Maximum Likelihood Rotated Factor Solution - Black Referent.....	123
Table 9. Multi-Group CFAs Assessing Construct Equivalence for Prejudice Scales ....	124
Table 10. Multi-Group CFAs Assessing Construct Equivalence for Motivation to Hide Prejudice Scales .....	125
Table 11. RCM Main Effects Models of Confounds Related to the Participants.....	126
Table 12. RCM Interaction Models of Confounds Related to the Participants .....	127
Table 13. Correlations amongst Study Variables .....	129
Table 14. Raters' Assessment of Applicants – Student Sample .....	130
Table 15. Predictors of Difficulty Parameters for Latent Class 1.....	131
Table 16. Predictors of Difficulty Parameters for Latent Class 2.....	132
Table 17. Predictors of Difficulty Parameters for Latent Class 3.....	133
Table 18. Predictors of Difficulty Parameters for Latent Class 4.....	134
Table 19. Predictors of Difficulty Parameters for Latent Class 5.....	135
Table 20. CFAs for Prejudice Scales, Adult Replication.....	136
Table 21. Multi-Group CFAs Assessing Construct Equivalence for Prejudice Scales, Adult Replication .....	137
Table 22. Maximum Likelihood One-Factor Solution, 15 Items.....	138
Table 23. CFAs for Prejudice Scales, Adult Extension .....	139
Table 24. Multi-Group CFAs Assessing Construct Equivalence for Prejudice Scales, Adult Extension .....	140
Table 25: External Motivation to Hide Prejudice, Adult Sample.....	141
Table 26. Ethnic Identity CFA.....	142
Table 27. Multi-Group CFAs Assessing Construct Equivalence for Ethnic Identity.....	143



Table 28. RCM Main Effects Models of Confounds Related to the Participants.....	144
Table 29. RCM Interaction Models of Confounds Related to the Participants .....	145
Table 30. Between-Participant Correlations, Adult Sample.....	148
Table 31. Raters' Assessment of Applicants – Adult Sample .....	151
Table 32. Predictors of Difficulty Parameters for Latent Class 1.....	152
Table 33. Predictors of Difficulty Parameters for Latent Class 2.....	153
Table 34. Predictors of Difficulty Parameters for Latent Class 3.....	154
Table 35. Predictors of Difficulty Parameters for Latent Class 4.....	155
Table 36. Predictors of Difficulty Parameters for Latent Class 5.....	156
Table 37. Estimated Marginal Means across Classes .....	157
Table 38. Estimated Marginal Means across Samples.....	158
Table 39. Chi-Square Tests of Demographic Differences in Latent Profiles (Student) .	160
Table 40. Chi-Square Tests of Demographic Differences in Latent Profiles (Adult) ....	161
Table 41. RCM Main Effects Models of Confounds Related to the Design .....	162
Table 42. RCM Interaction Models of Confounds Related to the Candidates .....	163

## List of Figures

Figure 1.1 Response Probabilities in a Polarized Scale .....	164
Figure 1.2 Response Probabilities Given Equivalent Usage of Scale Markers .....	164
Figure 1.3 Difficulty Parameters in Polarized Scale.....	164
Figure 1.3 Difficulty Parameters in Polarized Scale.....	165
Figure 2.1 Black Raters' Assessment of Applicants – Archival Sample.....	166
Figure 2.2 White Raters' Assessment of Applicants – Archival Sample .....	166
Figure 3. Screenshot of Rating Page.....	167
Figure 4.1 Scree Plot for Black-Referent Items, Student Sample (Reverse Included)...	168
Figure 4.2 Scree Plot for White-Referent Items, Student Sample (Reverse Included)...	169
Figure 5.1 Scree Plot for Black-Referent Items, Student Sample (Reverse Excluded)..	170
Figure 5.2 Scree Plot for White-Referent Items, Student Sample (Reverse Excluded) .	171
Figure 6.1 Black Raters' Assessment of Applicants – Student Sample .....	171
Figure 6.1 Black Raters' Assessment of Applicants – Student Sample .....	172
Figure 6.2 White Raters' Assessment of Applicants – Student Sample .....	172
Figure 7.1 Ratings of Candidates – Latent Class One .....	173
Figure 7.2 Ratings of Candidates – Latent Class Two.....	173
Figure 7.3 Ratings of Candidates – Latent Class Three.....	174
Figure 7.4 Ratings of Candidates – Latent Class Four .....	174
Figure 7.5 Ratings of Candidates – Latent Class Five.....	175
Figure 8.1 Scree Plot for Black-Referent Items, Adult Sample.....	176
Figure 8.2 Scree Plot for White-Referent Items, Adult Sample .....	177
Figure 9.1 Black Raters' Assessment of Applicants – Adult Sample.....	178
Figure 9.2 White Raters' Assessment of Applicants – Adult Sample .....	178
Figure 10.1 Ratings of Candidates – Latent Class One, Adult Sample .....	179
Figure 10.2 Ratings of Candidates – Latent Class Two, Adult Sample .....	179
Figure 10.3 Ratings of Candidates – Latent Class Three, Adult Sample .....	180
Figure 10.4 Ratings of Candidates – Latent Class Four, Adult Sample .....	180
Figure 10.5 Ratings of Candidates – Latent Class Five, Adult Sample.....	181

The duality of bias: Predictors of racially motivated Differential Test Functioning in  
interview evaluations

Racial discrimination in hiring and appraisal remains a salient concern in the workplace today (Huffcut & Roth, 1998; McKay & McDaniel, 2006; Roth, Bobko, McFarland, & Buster, 2008). However, despite years of research documenting the presence of discrimination in hiring, its exact nature is not known. Specifically, discrimination may result from raters *favoring* one group (e.g., pro-Caucasian) or discrimination may result from raters *penalizing* a different group (e.g., anti-black). Indeed, discrimination might result as a function of raters exhibiting both preferential and derogatory responding. However, while either type of responding may operate during decision-making, prior research suggests that both do not always occur simultaneously (Brewer, 1979; 1999; Brown, 2000), but rather that raters tend to favor individuals belonging to one group or penalize individuals belonging to another group at any given time.

In addition to a lack of specificity regarding the nature of the process underlying discrimination, the prior empirical literature has not clarified when differential responding stems from social and motivational factors and when it stems from individual difference factors, such as stereotypes and prejudice. If we better understood the process by which discrimination emerges, and what drives this process, researchers would be able to formulate a more consistent and targeted strategy toward investigating the complexities of discrimination. Moreover, understanding this process is critical in determining interventions that might be employed to reduce discrimination. Specifically,

strategic interventions to reduce discrimination would differ depending on whether an individual favors one group, or penalizes another.

The current study was designed to address these issues. Specifically, I first discuss the literature on racial discrimination in evaluative contexts. Then, I review how Item Response Theory (IRT) and Latent Class Mixture Modeling (LCMM) provide a platform for determining patterns in favoritism and derogation in rating scale usage across groups. Next, I will discuss a critical social factor that is relevant to discrimination: group membership and corresponding intergroup bias. I will then connect this social factor to anticipated differences in rating scale usage in organizational decision-making settings. That is, I hypothesize that in-group belonging will predict favoritism toward members in the in-group for high-status individuals (e.g. white raters), whereas in-group belonging will predict derogation against members in the out-group for low-status individuals (e.g. black raters). Then, I will discuss individual differences (e.g. stereotypes and prejudice) that also might drive differential rating scale usage. I will likewise relate these individual differences to expected behavior in organizational decision-making settings. Specifically, I hypothesize that prejudice predicts derogation against members of a sub-group, not favoritism toward members a sub-group. I then discuss how different combinations of prejudice and in-group belonging will result in different patterns of job candidate evaluation.

Finally, I outline three studies that assess different portions of the current theory. In the first study, I analyzed archival information of ratings of black and white candidates for an entry-level firefighting position to assess intergroup bias toward in-group members. In the second and third studies, I conducted an experiment including student

and adult participants to replicate and extend the findings from Study 1 in a more controlled environment. All three studies employ IRT analyses, and the two experimental studies also employ LCMM.

In sum, the current paper reviews the development and testing of a series of hypotheses regarding how patterns of job candidate evaluation are expected to result from different combinations of individual differences and social factors. Specifically, I aim to differentiate predictors of favoritism from derogation in an evaluative setting. Hypotheses will be tested in several studies investigating an area of inquiry in which discrimination is clearly prevalent: race and social interaction competence. Thus, I will next review prior work on racial stereotypes, prejudice, and discrimination in the workplace.

### **Racial Discrimination in Evaluations**

One of the most prolific topics of study in the prejudice and discrimination literatures pertains to racial bias. Extensive work on this topic has found that blacks are generally viewed in a more negative light than whites. In particular, the content of stereotypes of African-Americans includes assumptions of laziness (Brigham, 1971), insecurity (Butt & Signor, 1976), poverty, aggressiveness (Lepore & Brown, 1997), a lack of education, low intelligence (Devine, 1989), and low competence (Fiske, Cuddy, Glick, & Xu, 2002). Given the focus of these views on issues relevance to performance and ability, the stereotypes held against this group have profound implications on personnel selection concerns.

Indeed, there is a long history of work in racial discrimination in hiring (e.g. interview evaluations: Parsons & Liden, 1984; Pulakos, White, Oppler, & Borman, 1989) and performance appraisal (Mobley, 1982; Waldman & Avolio, 1991). Meta-analyses in

these areas consistently reveal black-white subgroup differences (Huffcut & Roth, 1998; McKay & McDaniel, 2006; Roth, Bobko, McFarland, & Buster, 2008). For example, a meta-analysis on discrimination in the structured interview shows that blacks are evaluated at about a quarter of a standard deviation below whites, which is a much smaller discrepancy than found in cognitive ability tests and low-structured interviews (Huffcut & Roth, 1998). Likewise, Goldstein, Yusko, Braverman, Smith, & Chung (1998) found that there were black-white subgroup differences on outcomes of assessment center exercises. Moreover, Goldstein et al. (1998) found that subgroup differences on these exercises were linked to subgroup differences in cognitive ability. However, whether differences in cognitive ability represent true intelligence differences, or if they are due to differences in socialization, culture, or other contaminating factors, was not determined.

In terms of assessment, multiple meta-analyses (Ford, Kraiger, & Schechtman, 1986; McKay & McDaniel, 2006) find that for most measures of performance (e.g. absenteeism) larger subgroup differences exist when the measurement employed is subjective. Moreover, research shows that blacks are rated lower than whites in terms of expected typical performance (DuBios, Sackett, Zedeck, & Fogli, 1993), and that supervisors provide lower job evaluations and have lower perceptions of promotability for African-American employees (Greenhaus, Parasuraman, & Wormley, 1990). Overall, it appears that subgroup differences are strongest for cognitive criteria relative to evaluations of social or interpersonal skills (Huffcut & Roth, 1998; McKay & McDaniel, 2006).

It is apparent that racial discrepancies in evaluations persist. The next concern, then, is whether these discrepancies exist due to preference for whites, derogation of blacks, or some combination of both. In this respect, the literature is not entirely clear. Specifically, some empirical findings seem to indicate that black-white discrepancies may be due, at least in part, to pro-white favoritism. For example, Bass & Turner (1973) found that there was a stronger relationship in managerial ratings of objective and subjective criteria for black ratees relative to whites. In other words, job-irrelevant information appears to increase performance evaluations for whites, which would indicate a pro-white preference. Similarly, Dovidio & Gaertner (2000) uncovered what appears to be favoritism for whites in evaluating ambiguous qualifications. Specifically, in this study, black and white targets were recommended equally for hire when provided qualifications were clearly very low or very high. However, when qualifications are ambiguous, around 70% of white targets were recommended for hire, while around 50% of black targets were recommended. Assuming that ambiguous information should lead to arbitrary decision-making, the expected percentage to forward should be around 50%. Given the strength of recommendation for ambiguously qualified white targets, pro-white favoritism seems to have entered into these evaluations.

While these studies seem to indicate that favoritism toward whites may affect evaluations, an overwhelming portion of the literature suggests just the opposite: that the differential responding in performance evaluations stems from derogation of African-American targets. Indeed, work by a number of scholars emphasizes that negative stereotypes are more frequently held against blacks rather than whites (e.g. Fiske et al., 2002). Correspondingly, research has successfully highlighted the role of negative

stereotypes with lower performance evaluations of African Americans (Baltes, Bauer, & Frensch, 2007), and worse recall for interview answers (Frazer & Wiersma, 2001). Moreover, Jussim, Coleman, & Lerch (1987) found that black applicants are rated in a more extreme manner, resulting in a broader range of responses to black rather than white applicants. In terms of distinguishing between different levels of ability, Hamner, Kim, Baird, & Bigoness (1974) found that high-performing black applicants were merely seen as average, and were rated as only slightly better than low-performing blacks. In this research, white targets were sorted in a more objective fashion—with low-performing whites rated low and high-performing whites rated high. Similarly, Mullins (1982) found that participants cannot distinguish between high and low performing black applicants, but do distinguish between low and high performing white applicants. A more recent study replicated these findings, and found additional support that these differences are exaggerated when blacks are evaluated for high-status jobs (King, Madera, Hebl, Knight, & Mendoza, 2006). Finally, heterogeneity of scoring for the same target was uncovered by Grove (1981), who discovered higher inter-rater agreement on ratings of white applicants than on ratings of black applicants.

Thus, prior research seems to indicate that individuals generally seem to respond negatively toward blacks, rather than exhibiting favoritism towards whites. However, there is some support for the opposite conclusion: that differential responding favors whites, but doesn't necessarily reflect derogation of blacks. One explanation for the disparity in these results is that such tendencies do not concern aggregate differences in evaluations across participants. Instead, differences in usage of the Likert scale occur within a given individual evaluator. As such, it is probable that some individuals exhibit



favoritism toward individuals belonging to a particular group, while others penalize individuals from a different group, and still others exhibit both favoritism and derogation. If, indeed, these tendencies reflect individual characteristics, then different patterns of responding should be apparent across individuals in evaluative contexts. Moreover, these patterns should be associated with relevant individual difference characteristics or social factors. Next, I review literature on potential social factors that may drive differences in responding to members of different sub-groups in an evaluative context.

### **Social Influences of Discrimination**

A large body of research has been conducted on social influences of discrimination. In particular, discrimination can arise in part from identification with a given group. Specifically, according to Self-Categorization (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987) and Social Identity (Hogg & Abrams, 1988; Tajfel & Turner, 1979) theories, individuals seek to make social comparisons of themselves with others in order to build self-esteem and reduce uncertainty. While in-groups may be determined based off of some pre-existing characteristic, such as race or gender, they can also be formed rapidly based on some salient attribute of a given situation. For example, in prior research, in-group membership has been successfully primed by providing participants differently-colored booklets (Vanbeselaere, 1993), or informing participants that they belonged to some fictional group, such as a “Klee” or “Kandinsky” group (Peterson & Blank, 2003).

The process of social comparison results in classifying others as part of “in-groups” or “out-groups”. Moreover, identification with in-groups prompts behavior aimed at maintaining distinctiveness between in-groups and out-groups (e.g. Scheepers,

Spears, Doosje, & Manstead, 2006). Such behavior can lead to discrimination of out-group individuals. Specifically, this phenomenon is known as the intergroup bias.

Intergroup bias is conceptualized broadly as the systematic attitudinal and perceptual biases that favor members of some defined “in-group” over some defined “out-group”, and is strongly associated with intergroup competition (Brewer, 1979; Hewstone et al., 2002; Lipponen & Leskinen, 2006). Thus, the phenomenon of intergroup bias suggests that raters should evaluate candidates differently depending on the match between rater and applicant race. That is, I hypothesize that raters will evaluate candidates of their own race more leniently than candidates of another race.

*Hypothesis 1: There will be a significant interaction between the race of the rater and race of the applicant on usage of Likert markers. Specifically, raters will be more likely to be more lenient toward same-race applicants rather than other-race applicants.*

In order to assess this hypothesis, it is imperative to address how differential rating scale usage will be conceptualized and measured in the current research. Moreover, it is necessary to specify how “lenience” will be operationalized. To this end, I next discuss how IRT and LCMM can be employed to better understand differential scale usage across groups.

### **Item Response Theory and Differential Test Functioning**

IRT is forwarded as a particularly powerful tool that can be used to assess favoritism versus derogation in evaluations. IRT is a theoretical framework developed to better understand ability and error. Developed in the field of educational testing, initial IRT models sought to predict “correct” and “incorrect” responses to questions. In contrast to Classical Test Theory (CTT), IRT does not assume that the observed scores on such

questions are a function of only true score and random error. Instead, IRT allows for the existence of other systematic influences on the observed score, such as item difficulty (included in the Rasch model), item discrimination between individuals at a particular skill level (the 2-parameter logistic model), and guessing (the 3 parameter logistic model).

While IRT models were originally designed to assess binary responses, they have also been adapted to address responses to questions that have more than 2 outcomes. Indeed, a number of so-called polytomous models have been developed to better understand how people respond to a range of options. For example, these methods can be applied to assess the apparent underlying psychological distances between markers on Likert scales in terms of “difficulty” of ascending from one marker to the next. These models vary on a variety of assumptions, including whether or not steps to each successive marker must be of equal or ascending difficulty levels.

Although psychological researchers have not widely applied polytomous IRT models to the systematic study of differential test functioning, the application thereof is fairly straightforward. Specifically, two relevant concepts that have emerged primarily within the dichotomous IRT literature are the concepts of Differential Item Functioning (DIF) and Differential Test Functioning (DTF). DIF occurs when items are differentially difficult for individuals in one focal group relative to another (Meulders & Xie, 2004). Similarly, DTF occurs when differences between focal groups in item difficulty result in differences across focal groups in test characteristics (Meulders & Xie, 2004). Thus, DIF and DTF represent interactions between focal groups and item or test functioning in understanding the relative difficulty of items or tests, respectively. In the current study,

DTF will be employed to assess differences in rating scale usage across referent groups that might result in discriminatory outcomes. One particular polytomous IRT model is especially well-suited for testing differences in distances between Likert markers on a scale—the Partial Credit Model (PCM; Masters, 1982).

While the PCM model was initially generated for use in achievement tests where there are multiple steps, it is also useful in assessing attitude scale responses (Masters & Wright, 1996). The PCM could be fit to the data using the following equation to estimate step difficulty parameters:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^x (\theta_n - \delta_{ij})\right]}{\sum_{x=0}^{m_i} \left[\exp\left[\sum_{j=0}^x (\theta_n - \delta_{ij})\right]\right]} \quad \text{Equation 1.1}$$

In this equation,  $\theta_n$  is the ability parameter for a given individual,  $n$ , on a latent continuum,  $m_i$  is the maximum score (e.g. “5” or “7”) for a particular item,  $i$ , and  $\delta_{ij}$  is the difficulty step for the  $j^{\text{th}}$  threshold between two response categories for a particular item (e.g. between “1” and “2”). Thus, scores for a given individual on item  $i$  follow random variables that can take on any integer value from  $x_{ni} = 0, \dots, m_i$ . Notably, the difficulty parameters across thresholds are calculated such that  $\sum \delta_{ij} = 0$ . That is, these parameters sum to zero. Consequently, this equation specifies that the probability that the  $i^{\text{th}}$  item will take on a score of  $x$  for a given individual,  $n$ , is related to the difficulty parameters for the thresholds between item categories ( $d_{ij}$ ) and the ability of that individual ( $\theta_n$ ).

When the PCM is employed to assess candidate ratings,  $\theta_n$  can be conceptualized as the ability level of a given candidate. Similarly,  $\delta_{ij}$  would capture how “difficult” it is for a specific rater to move from a rating of a “1” to a rating of a “2” (or “2” to “3”, etc.)

for a given candidate. More specifically, the first step difficulty parameter specifies how high an individual's ability level has to be (relative to any other step in the scale) in order to be rated a "2" rather than a "1". For example, if  $\delta_{a1}$  (the difficulty parameter for the first step) is -0.30, then candidates with ability levels greater than -0.30 will be classified in the second category (as a "2"), whereas candidates with ability levels less than -0.30 will be classified in the first category (as a "1"). Similarly, the second step difficulty parameter specifies how high an individual's skill level has to be, relative to any other step in the scale, in order to be rated a "3", rather than a "2". In sum, then, this equation specifies that the probability that a target is evaluated at a particular skill level is a function of their ability and of raters' usage of points on the Likert scale.

One of the key issues for the current series of studies is to understand how to interpret the PCM difficulty parameters. While some polytomous IRT models (e.g. the graded response model; Samejima, 1997) assume that successive difficulty parameters must necessarily increase in difficulty, the PCM (Masters, 1982) has no such assumption (Embretson & Reise, 2000). Indeed, under the PCM, it is possible to both have successive steps which are easier (i.e., more negative difficulty parameter) to attain, or exceed, than the prior steps, or to have steps of equivalent difficulty. Figures 1.1 and 1.2 provide approximate graphical depictions for such items.

Figures 1.1 and 1.2 depict curves representing approximate probabilities of receiving each rating, from 1 to 5, on a given item. Difficulty parameters in this figure are illustrated via the intersection of curves. At these intersections, a given individual is equally likely to be classified into either of the adjacent categories (e.g. equally likely to be classified as a "1", or as a "2"). Before each intersection, individuals are more likely to

be classified into the prior category (“1”), and after each intersection, individuals are more likely to be classified into the latter category (“2”). In instances where successive steps are easier to exceed relative to prior steps, successive thresholds (such as between 2 and 3) may nearly overlap with prior thresholds (such as between 1 and 2).

Figure 1.1 displays approximate probability curves for an item where successive steps are “easier” than prior steps. In this figure, it is apparent that responses to that item tend to be either “1” or “5”. In other words, ratings on this item are polarized.

Consequently, it is difficult for a rater to move past an evaluation of “1” for a given target. However, if this rater does move past an evaluation of “1”, evaluations of “2” and “3” are passed entirely, as the difficulty parameters associated with these thresholds are much lower than the difficulty parameters associated with the first threshold. In instances where steps are of equivalent difficulty, the “distance” between thresholds is also equivalent. In contrast to Figure 1.1, Figure 1.2 displays approximate probability curves for an item where successive steps are equivalent in difficulty to prior steps.

Finally, Figure 1.3 provides a direct depiction of the difficulty parameters associated with an item for which successive steps are easier to attain than prior steps, rather than their associated probabilities. Thus, Figure 1.3 provides another way of looking at responses to the item depicted in Figure 1.1. Specifically, this graph reveals that the difficulty parameter for the first threshold (between “1” and “2”) is very high, whereas the difficulty parameters for each of the following thresholds (between “2” and “3”, “3” and “4”, and “4” and “5”) are all lower than this first difficulty parameter. The implication of this property is that categories associated with easier difficulty parameter are essentially not used by raters. Thus, even though the item shown in Figures 1.1 and

1.2 follows a 5 point scale, this scale really functioned as a 2 point scale because three of the difficulty parameters were lower than the parameters that immediately preceded them.

That is, as previously discussed, discrimination may emerge either from favoritism toward individuals in a given group, or from derogation against individuals in another group. Thus, DTF might emerge due to differential rating scale usage in favor of a given group, or from differential rating scale usage against another group. As previously mentioned, difficulty parameters capture the ease or difficulty associated with a particular rater's ascension up the Likert scale. Additionally, it is important to note that step difficulty parameters for a given rater will sum to 0. Thus, some difficulty parameters will be positive, and others negative, for the same rater. Those difficulty parameters which are positive indicate steps that are "difficult" for candidates to pass. In other words, candidates would need a higher skill level in order to be evaluated using the higher number. Those difficulty parameters which are negative indicate steps that are "easy" for candidates to pass. In other words, candidates would need a lower ability (relatively speaking) in order to be evaluated using the higher number.

Given that step difficulty parameters in the PCM capture relative difficulty, a key issue in determining whether or not DTF is operating focuses on step difficulty parameter *magnitude* differences between individuals for different "groups". As discussed, IRT difficulty parameters are estimated such that the difficulty parameters across all steps for an item sum to zero. However, for a particular item, the difficulty parameters may all be very close to zero (similar relative step difficulty), or may show great variation around zero (large variance in relative step difficulty). Hence, it seems straightforward to conclude that a rater exhibiting DTF would exhibit differential variation of the difficulty

parameter across groups whereas a rater who does not exhibit DTF would exhibit similar variation of the difficulty parameter across both groups.

However, examining the overall variance of the difficulty parameters across racial groups for each rater may yield the false impression that a rater is using the scale in the same way for members of two groups. For example, a rater could evaluate candidates severely when assessing whether black applicants are minimally competent on some dimension (i.e., large positive difficulty parameters for the low end of scale), but may be extremely lenient if s/he perceives that black applicants exceed the minimal competence cut-off on the dimension (i.e., large negative difficulty parameters for the upper end of scale). That same rater may also be extremely lenient when assessing whether white applicants are minimally competent (i.e., large negative difficulty parameters for the low end of scale), but may be harsher when evaluating white applicants in the competent range of the scale (i.e. large positive difficulty parameters for the high end of the scale).

In this scenario, this rater is clearly using the scale differently as a function of applicant race, yet the variance of the difficulty parameters that s/he is exhibiting throughout the entire scale is equivalent for the two groups (because s/he is differentially severe or lenient depending upon the applicant race and the level of the scale). Thus, a more nuanced perspective is required to assess the existence and direction of DTF. That is, individual steps must be compared across groups in order to understand if—and where—DTF might be occurring within the scale. In particular, comparisons across groups of the difficulty parameters associated with the first two steps (assuming a 5-point scale) should provide evidence for the existence and direction of DTF.



Specifically, if the first difficulty parameters are equally “hard” for both groups, then DTF is present. However, if the first difficulty parameters are relatively “hard” for candidates in one group, but not for another, then the rater in question over-uses “1’s” or “2’s” for candidates from one group relative to another. Conversely, if the first difficulty parameters are relatively “easy” for candidates in one group, but not for another, then the rater in question is under-using “1’s” or “2’s” when rating candidates in that group relative to another. Thus, assessing differences in the first two parameters across groups can reveal the existence and direction of DTF.

In sum, the existence and nature of DTF can be detected by examining individual step difficulty parameters across groups. If there are no apparent differences across groups at each step, then no DTF is present. If step difficulty parameters for one group do not vary greatly from zero, and step difficulty parameters for another group do vary greatly around zero, then a rater is exhibiting DTF in favor of individuals in one group, or against individuals in the other. Finally, if both average absolute step difficulty parameters vary greatly from zero, both DTF-against (derogation) and DTF-for (favoritism) may be in operation.

Finally, latent class mixture modeling (LCMM) can be employed to determine distinct sub-populations who exhibit different patterns of responding. That is, most statistical analyses in organization research are conducted under the assumption that the researcher is sampling from one specific population. LCMM, on the other hand, is an analysis technique that allows for the estimation of different sub-populations based on patterns of responding (Wang & Hanges, 2011). That is, this procedure identifies latent groups of participants. Across these groups, response patterns vary, while within these

groups, there is lower variation in terms of patterns of responses. For the current study, I ran LCMMs on raters' step difficulty parameters for black and white candidates.

Specifically, I sought to categorize raters based on their differential usage of the scale points for black and white job candidates. Membership in these latent classes can then be statistically predicted by person characteristics.

In sum, differential scale usage is expected to manifest as DTF across certain subgroups, whereas no-DTF manifests as equivalent usage of Likert scale markers across candidates from different subgroups. As such, a combination of IRT and LCMM provides a methodology that lends itself well to assessing differential scale usage in interview evaluations. Next, I apply this operationalization of DTF in assessing hypothesized racial differences in archival interview evaluations.

## **Study 1 Method**

### **Participants**

The participants of the present study were raters of a firefighter selection interview process developed and used during 2007. A total of 19 raters evaluated the responses of 318 entry level firefighter applicants. The raters consisted of nine blacks, nine whites and one "other". Raters were experienced Captains and Lieutenant firefighters. Raters were recruited from throughout the continental United States.

### **Stimuli**

As discussed, archival data of interview ratings for 318 black and white firefighter applicants was employed for this initial study. Entry level firefighter applicants were provided with the five situational judgment questions directly prior to their interview, and given 25 minutes to prepare their answers (which amounts to roughly 5 minutes per

question). They were allowed to take notes to help them formulate their responses to each question and were allowed to take their notes into the actual interview setting.

Immediately following the allowed preparation time, applicants were taken into solitary rooms. Each room was equipped with a computer and a video camera. Questions were presented both visually and audibly, and the applicant was given four minutes to verbally respond to each.

The situational judgment questions employed in the interview were developed following a content valid procedure. The first situational judgment question asked the applicant how he or she would respond in a situation where another firefighter was not pitching in to do his or her fair share. The second situational judgment question asked the applicant how he or she would react in a group work situation where his or her colleagues were struggling with their assigned tasks. The third situational judgment question asked the applicant how he or she would prepare to take an Emergency Medical Technician exam. The fourth situational judgment question asked the applicant how he or she would respond to a civilian interruption at the firehouse at 2 a.m. Finally, the fifth situational judgment question asked the applicant how he or she would deal with the emotions resulting from a near-death experience. Applicants responded to these questions verbally.

### **Ratings and Frame-of-Reference Training**

Applicant responses were rated on 5 dimensions: interpersonal skills, team-orientation, learning-orientation, customer-service orientation, and stress management. The rating scales ranged from 1 (unacceptable) to 5 (outstanding). Behavioral benchmarks were provided for each rating scale to facilitate the raters understanding of

the meaning of the scale anchors. All raters received a two day frame-of-reference training before rating any applicants.

Frame-of-reference training was provided to attempt to synchronize raters' frames of reference and to minimize personal biases in responding (Bernardin & Buckey, 1981). In frame-of-reference training, assessors are educated on desirable job-related behaviors, provided with opportunities to practice rating candidates, and given constructive feedback on rating accuracy (Pulakos, 1986). Indeed, current research suggests that frame-of-reference training increases rating accuracy and minimizes biases (Woehr & Huffcutt, 1994) as well as increases consistency in assessor ratings (Schleicher, Day, Mayes, & Riggio, 2002). Two raters provided evaluations for each candidate. The questions asked of candidates and the instruments used for evaluation are available in Appendix A.

As discussed, rater training for the structured interview was conducted. This training lasted a full day and included a description of the test development process, general interviewer rater training (e.g., how to avoid rating errors, taking notes, etc.), discussions regarding each interview question and associated benchmarks, an explanation of the rating process, practice sessions rating actual interview questions, and so forth.

### **Procedure**

Each rater worked with between 3-5 other raters over the course of the assessment. Consequently, a total 45 rotating pairs of black and white raters were formed. On average, then, each pair rated around seven candidates. Applicant responses were video-taped and raters were provided these tapes to evaluate each applicant. Each applicant was rated by two raters (i.e., one student and one firefighter). The rater team

worked together for one day and then team members were randomly assigned to a new rater team the following day.

### Study 1 Results

I first assessed whether there was an overall “effectiveness” construct among the five interview rating dimensions. This was done by conducting a confirmatory factor analysis in MPLUS. This analysis shows that indeed, all five ratings load on to one factor representing overall effectiveness ( $\chi^2(5) = 21.56, p < 0.05$ ; CFI = 0.98; RMSEA = 0.07 (CI: 0.04 – 0.11); SRMR = 0.03). Since these items are all tapping into the same latent construct of overall effectiveness, I did not investigate individual item differences in the following analyses. Instead, I estimated step difficulties on each item given that they load on a single effectiveness latent construct.

Next, responses were analyzed using using Item Response Theory (IRT) to assess differences in step difficulties across raters. Specifically, data was fit to a series of partial credit models (PCMs) using ConQUEST. For the first hypothesis, I performed a series of PCM analyses. Specifically, I first imposed a PCM model that predicts step difficulty as a function of differences between raters. That is:

$$\delta_{ih} = w_1 rater_{k\bullet} + w_2 step + w_2 rater_{k\bullet} * step \quad \text{Equation 2.1}$$

This model allows for differences in difficulty steps between the Likert response categories. This model assumes that applicant race does not affect rater evaluations of the applicants, but that raters differ in difficulty between Likert scale markers. Indeed, this analysis reveals that rater characteristics clearly impact step difficulty parameters on average ( $\chi^2(17) = 572.08, p < 0.05$ ). I next imposed a second PCM model on the data.

This second model included applicant race as an additive factor. Specifically, this model is:

$$\delta_{ih} = w_1 rater_{k\bullet} + w_2 step + w_3 rater_{k\bullet} * step + w_4 race_{\bullet l} + w_5 race * step \quad \text{Equation 2.2}$$

The second model assumes that there is an overall effect for applicant race that is consistent across all raters. According to this analysis, rater ( $\chi^2(17) = 785.39$ ,  $p < 0.05$ ) predicts difficulty parameters on average. Race of applicant ( $\chi^2(1) = 4.81$ ,  $p < 0.05$ ) also appears to predict difficulty parameters overall, such that whites have lower step difficulty parameters on average than blacks ( $w_4 = -0.06$ ).

The third model adds a rater by race interaction to reflect the possibility that some raters exhibit different DTF relative to others. Specifically:

$$\delta_{ih} = w_1 rater_{k\bullet} + w_2 step + w_3 rater_{k\bullet} * step + w_4 race_{\bullet l} + w_5 race * step + w_6 rater_{k\bullet} * race_{\bullet l} + w_7 rater_{k\bullet} * race_{\bullet l} * step \quad \text{Equation 2.3}$$

The difference between this model and the two aforementioned models is that this model assumes that raters are differentially sensitive to applicant race. Analysis of this model reveals that there is a rater by applicant race interaction in step difficulty parameters ( $\chi^2(17) = 2,250.04$ ,  $p < 0.05$ ). While suggestive, this analysis didn't directly test Hypothesis 1 which specified a particular direction to this interaction. That is, Hypothesis 1 predicted that raters should be more lenient (i.e., lower step difficulty parameters) when rating same-race relative to different-race applicants. Thus, to analyze this hypothesis, I ran another PCM in ConQUEST specifying that rater race interacts with applicant race to predict step difficulty parameters. Specifically:

$$\delta_{ih} = w_1 raterrace_{k\bullet} + w_2 step + w_3 raterrace_{k\bullet} * step + w_4 race_{\bullet l} + w_5 race * step + w_6 raterrace_{k\bullet} * race_{\bullet l} + w_7 raterrace_{k\bullet} * race_{\bullet l} * step \quad \text{Equation 3.1}$$

This equation builds on the prior equations in that it explicitly specifies the relevant characteristics of the rater (race) as interacting with applicant race in predicting step difficulty parameters. Evidence for Hypothesis 1 would be provided by a significant rater by race interaction, provided that the direction of the interaction is consistent with the aforementioned hypothesis. Indeed, there is a significant interaction between applicant race and rater race ( $\chi^2(1) = 18.69$ ,  $p < 0.05$ ) in predicting average difficulty parameters. Specifically, black raters appear to use the rating scale more consistently on average when rating black applicants than white raters (Average deviation for black raters = 0.13; Average deviation for white raters = 0.21). Similarly, white raters appear to use the rating scale more consistently when rating white applicants than white raters (Average deviation for black raters = 0.13; Average deviation for white raters = 0.08). However, as previously discussed, the omnibus test of step difficulty parameters may not provide enough information to assess DTF in evaluations.

Thus, I next assessed differences between each step difficulty parameter across black and white applicants. Specifically, I employed the estimated standard errors for steps to assess whether the step difficulty parameters overlapped across applicant race within raters. That is, if the step difficulty parameters do not overlap at particular thresholds, based on the standard errors, DTF in individual steps can be detected. Table 1 provides the difficulty parameters and standard errors for the Black and White raters as a function of applicant race. Figure 2 provides a pictorial representation of the same information.

An analysis of black raters' assessment of white and black applicants reveals no statistically significant differences in difficulties for specific steps, based on the estimated

standard errors. Specifically, black raters did not have more lenient difficulty parameters for black applicants as compared to white applicants at each step (Step 1:  $t(9) = -0.49$ ,  $p > 0.10$ ; Step 2:  $t(9) = -0.66$ ,  $p > 0.10$ ). The same was true for white raters (Step 1:  $t(8) = 1.24$ ,  $p > 0.10$ ; Step 2:  $t(8) = -1.99$ ,  $p > 0.10$ ). As previously discussed, support for DTF-for would be declared if significant differences were detected in specific steps. Thus, the present study did not provide support for the hypothesis.

### **Study 1 Discussion**

Study 1 provided initial assessment of the usefulness of IRT in understanding DTF based on applicant race. Specifically, this study provided evidence demonstrating that the IRT methodology could be used to assess DTF in real work settings. That is, this study showed that the IRT model successfully captured how black and white raters responded to applicants of the same and different race. The analyses revealed that while the specific step at which the DTF occurred could not be identified, there was evidence that raters responded differently to applicants as a function of the match between their races. However, given that the more detailed step analysis failed to find significant effects, I concluded that this study failed to support Hypothesis 1. That is, I did not support the hypothesis that raters would be more lenient toward same-race applicants.

While the present study was useful in demonstrating that the IRT approach could be used in a real world context, it is important to point out limitations that hinder its interpretation. First, as previously discussed, the raters in the current study were provided with extensive frame-of-reference training. Since frame-of-reference training has been found to reduce personal biases in prior work (Woehr & Huffcutt, 1994), the lack of significant findings in the current study may reflect the efficacy of the training



rather than the validity of the current theory. Consequently, it is important to assess naïve raters for a more rigorous test of how DTF-for and DTF-against might manifest in interview evaluations.

Moreover, while I was able to investigate differences between raters' evaluations of black and white applicants, the difficulty parameters across races at each step were in the same direction. As such, the cause of the differences in ratings is not immediately clear. Thus, it is difficult to identify if any apparent differences were evidence of favoritism toward in-group members (less "hard" on same-race) or derogation of out-group members (more "hard" on other-race). To disentangle these phenomena, it is important to assess patterns in the responses of different sub-populations of individuals.

Additionally, while one strength of the current study was the number of applicants evaluated, the current theory and, indeed, the IRT methodology itself, is more focused on characteristics of the *rater*. This particular property has two implications on the ability of the current study to provide a powerful test of the current theory. First, the statistical power of this study is more a function of the number of raters than the number of applicants. Thus, the sample size of 19 raters is too small to adequately test Hypothesis 1 at the step level of analysis. Moreover, since the primary thrust of the IRT approach to DTF is to focus on the characteristics of the rater, hypotheses regarding characteristics of raters could not be tested in this field sample due to the low number of raters.

As such, the work assessing Hypothesis 1 in the field, while promising, provides a limited assessment of DTF in interview evaluations. Thus, I will next test a series of hypotheses in a laboratory setting to further explore DTF under more controlled

conditions. Additionally, I will be able to assess correlates of response patterns, and meaningfully connect them to DTF, in the following studies.

## **Study 2**

As discussed, Study 1 provides initial support for the utility of using IRT to capture DTF in an interview context. However, this study did not allow for a thorough construct validation of the current operationalization of DTF (i.e., the direction of DTF) and its drivers. Two experimental studies will thus be conducted in order to assess the direction of DTF, and to better understand what drives such threshold differences. To this end, I will first revisit the intergroup bias literature and discuss how in-group belonging may influence differences in responding to white and black applicants.

### **Intergroup Bias**

As previously discussed, the literature on intergroup bias suggests that raters should evaluate candidates differently depending on the match between rater and applicant race. However, the way in which the intergroup bias manifests may not be apparent based on this inference alone. Fortunately, while many studies on intergroup bias sought only to show that individuals treat in-group and out-group members differently, some work has been done to examine more thoroughly why such disparities emerge. Specifically, researchers on intergroup bias have begun to disentangle whether the disparate treatment of in- and out-group members reflects favoritism toward the in-group or derogation of the out-group.

The empirical literature on intergroup bias has generally found that favoritism of one group over another is a function of a person's positive orientation toward his/her own in-group as opposed to that person punishing members of the out-group (Brewer, 1979;

1999; Brown, 2000; Hewstone et al., 2002). That is, intergroup bias tends to stem from positive feelings toward the in-group, rather than negative feelings toward the out-group (Brewer, 2007; Hewstone et al., 2002). Hewstone et al. (2002) argue that distinct areas of research in intergroup bias support such a conclusion. First, positive evaluations of in-group members arise spontaneously, and are stronger than negative evaluations of out-group members (Otten & Wentura, 1999; Perdue et al., 1990). Second, the form of prejudice observed in most intergroup research is not traditionally hostile, but rather is characterized by fewer positive expressions toward the out-group as compared to the in-group (Dovidio & Gaertner, 2000; Stangor et al., 1991). In other words, it has been suggested that in-group favoritism is universal, whereas out-group derogation is more contextually contingent (Brewer, 2007).

The literature on intergroup bias thus implies that disparate treatment of in- and out-group members largely stems from same-race favoritism as opposed to different-race derogation. For example, prior studies on discrimination have found that whites raters are less lenient toward black defendants than black raters (Abwender & Hough, 2001) and that individuals are generally more lenient toward job applicants when the race of the rater and job applicant matched (Chatman & von Hippel, 2001). Extending the conclusions drawn from the intergroup bias literature to aid in the interpretation of such studies, it would seem that such discrimination may be a function of in-group favoritism as opposed to out-group punishment. Thus, the intergroup bias literature suggests that disparate treatment of in- and out-group members is driven by same-race preference rather than other-race derogation.

While such a prediction may appear to be useful in explaining favoritism toward in-group individuals in a given group, some research suggests that status may play a role in the manifestation of intergroup bias. For example, some research indicates that low-status groups exhibit negative (e.g. against the out-group) forms of intergroup bias than high-status groups (Scheepers et al., 2006), especially when status differentials are seen as unstable and/or justified (e.g. Ellemers, Wilke, & van Knippenberg, 1993). Conversely, members of low-status groups may exhibit out-group favoritism—but only when they feel that status differences are justified (Jost & Burgess, 2000).

Indeed, whites are considered to have a higher social status than blacks (Chattopadhyay, 1999; Tsui, Egan, & O'Reilly, 1992). However, this perceived status difference is not likely to be seen as justified. Thus, it is likely that in-group belonging predicts derogation of out-group members for black individuals, whereas in-group belonging should predict favoritism toward in-group members for white individuals.

The intergroup bias literature provides a strong rationale for favoritism toward the in-group in white individuals, but does not fully explain derogation against the out-group. A different framework might be necessary to explain why derogation of blacks would occur. Indeed, while intergroup bias is generally the result of positive evaluations of the in-group, some research has found that such bias also occasionally emerges due to derogation of the out-group. For example, when individuals associate stronger emotions with out-groups (Brewer, 2001) they may exhibit intergroup bias against out-group members. Such emotions may be prompted, for example, by apparent threat from the out-group (Hewstone et al., 2002), or from individual differences, such as prejudice against out-group members (Hewstone et al., 2002). Consequently, certain individual difference

factors, such as prejudice, are expected to influence derogation of individuals in a given group.

### **Other Individual Differences that Influence Differential Responding**

As discussed, individual differences can contribute to differential responding. Specifically, individuals form cognitive expectations of others based on the classification of these others into groups. That is, individuals hold *stereotypes* about others based on others belonging to different demographic categories, such as race (Stangor, 2009). When stereotypes are negative, individuals sometimes experience affective or attitudinal negative reactions to others—that is, individuals may be *prejudiced* against others (Stangor, 2009). While both stereotypes and prejudice are thought to have a social component, the extent to which they are endorsed varies across individuals (Schneider, 2004). As such, these individual differences likely influence judgments of individuals belonging to different categories.

While both stereotypes and prejudice affect perceptions of individuals in different subgroups, prejudice may be more proximal in its effects on discrimination than stereotypes. That is, stereotypes generally appear to impact prejudice (Schneider, 2004), which then impacts discrimination (Schutz & Six, 1996). Consistent with this proposition, a meta-analytic review of the literature on the impact of stereotypes and prejudice on discrimination reveals that prejudice tends to exhibit a stronger effect on discrimination than stereotypes (Talaska, Fiske, & Chaiken, 2008). Thus, for the purposes of the current study, I will investigate the role of prejudice in differential responding to white and black targets.

Prior research reveals a strong connection between prejudice and discrimination in an organizational context. For example, empirical evidence suggests that people who are prejudiced against blacks actively discriminate against blacks in hiring decisions (Ziegert & Hanges, 2005). Likewise, individuals who endorse negative stereotypes about blacks have been found to rate black applicants lower than those who do not strongly endorse such stereotypes (Baltes, Bauer, & Frensch, 2007). Since these racial stereotypes are negative and focus on African-Americans rather than whites, many raters might also exhibit intergroup bias against black applicants (i.e., DTF-against).

While prejudice may be a driver of differential responding, endorsement of prejudicial beliefs or stereotypes is not necessarily directed only toward out-group members. In other words, some individuals may display stereotypical or prejudicial beliefs toward members of their own in-group. That is, while whites may be prejudiced against blacks, it is possible that some black individuals may also have similar prejudices about their own group. For example, Clark & Clark (1947) found that the majority of black school children of that time choose to play with white dolls (over black dolls), due to the prevalence of anti-black prejudice in the broader US culture.

Moreover, prejudicial beliefs do not only exist with respect to assessments of black individuals. That is, people may exhibit prejudice against Arabs (Echebarria-Echabe & Guede, 2007), Asians (Lin, Kwan, Cheung, & Fiske, 2005), or even whites (Johnson & Lecci, 2003). Consequently, it is conceivable that individuals may exhibit prejudice against whites or blacks, regardless of their own racial in-group.

This review suggests that discrimination in an interview context is complex. That is, favoritism may be exhibited toward members of a given in-group, whereas derogation

may be directed toward members of a group against which negative prejudicial beliefs are held. Moreover, prejudicial beliefs may not apply only to out-group members. That is, some individuals may be prejudiced against their own apparent in-group. As such, both black and white individuals are likely to vary on prejudicial beliefs. Consequently, I expect that distinct patterns of ratings will emerge such that black and white individuals will display difference combinations of favoritism and derogation with respect to ratings of interview candidates. Thus, raters are hypothesized to differ in terms of the patterns of DTF that they exhibit. Five patterns are expected to emerge in the data. Specifically, individuals may exhibit (a) no DTF, (b) DTF against blacks but not in favor of whites, (c) DTF against blacks and in favor of whites, (d) DTF in favor of whites but not in favor of blacks, or (e) DTF against whites, but no DTF toward blacks. Thus, I hypothesize:

*Hypothesis 2: Five patterns of DTF will result in the data: a) no DTF, b) DTF against blacks and DTF in favor of whites, c) DTF against blacks but not toward whites, d) DTF in favor of whites but not toward blacks, and e) no DTF toward blacks but DTF against whites..*

Additionally, drawing from the prior literature, I hypothesize that these five combinations should be distinguished by similarity in applicant-rater race and prejudice. Specifically, DTF-for should generally be driven by in-group belonging (applicant-rater race similarity) for white raters, whereas DTF-against should generally be driven by prejudicial beliefs for white raters and in-group belonging (applicant-rater race similarity) for black raters. Corresponding, I hypothesize that:

*Hypothesis 3: Rater race and prejudice will predict patterns of DTF, such that:*

- a. *Individuals with high prejudice against whites will exhibit DTF against white applicants*
- b. *Individuals with high prejudice against blacks will exhibit DTF against black applicants*
- c. *Black individuals will exhibit DTF against white applicants*
- d. *White individuals will exhibit DTF in favor of white applicants*

## **Study 2 Method**

### **Participants**

Participants were 234 students, recruited through the University of Maryland SONA systems website. Although courses offered in the African American Studies and Sociology departments were approached for recruitment purposes, no additional subjects were acquired through these means.

The student sample was 28.6% black ( $n = 67$ ) and 71.4% white ( $n = 167$ ). Additionally, the student sample was 67.9% female ( $n = 159$ ). Psychology was the most well-represented major, with 44% of participants ( $n = 103$ ). Other participants were either undeclared (12.4%,  $n = 29$ ) or had another major (43.6%,  $n = 102$ ). No other major was represented by more than 5% of the total sample. Participant GPAs ranged from 1.8 to 4 (mean = 3.34, stdev = 0.43), and participant ages ranged from 18 to 26 (mean = 19.42, stdev = 1.39). Most participants were sophomores (34.9%,  $n = 81$ ), followed by freshmen (29.5%,  $n = 69$ ), juniors (20.5%,  $n = 48$ ), and seniors (14.5%,  $n = 34$ ). The student sample is 54.7% Christian ( $n = 128$ ), 23.5% Jewish ( $n = 55$ ), 6.8% Agnostic ( $n = 16$ ), and 5.1% Atheist ( $n = 12$ ). No other religious group was represented by at least 5% of the sample. Finally, 60.3% of the participants were democrats ( $n = 141$ ), 18.8% of the



participants were independent ( $n = 44$ ), and 13.2% were republican ( $n = 31$ ). No other political affiliation was represented by at least 5% of the sample.

### **Stimuli**

Archival videos of fifteen of the interview candidates assessed in Study 1 were obtained for use in the current study. Specifically, in order to maintain consistency between the archival and experimental studies, it was imperative to use actual candidate responses. These interview candidates were selected on the basis of quality of their responses, such that each of their five scores was relatively consistent. That is, five of these candidates were generally rated high, five were generally rated in the middle, and five were generally rated low. Additionally, in order to account for potential differences between gender and races in interview content, video responses were also selected on the basis of gender (male) and race (7 black, 8 white).

Pilot testing was conducted on the original videos to see if candidates had identifiable accents that may confound the results of the current study. Five undergraduate student raters who were blind to the study's hypotheses assessed the accents of the interview candidates. Specifically, raters were asked to rate the accents of each of these candidates on a scale of 1 to 5, with 1 being "no accent at all" and 5 being "heavy accent". Inter-rater reliability of these five raters was assessed using ICC1 and ICC2. The ICC1 was 0.77, indicating acceptable agreement between raters, and the ICC2 was 0.94, indicating that the average rating across all five raters was reliable.

Average accent ratings for the fifteen candidates ranged from 1 to 4.6. Unfortunately, there appeared to be a relationship between level of accent and evaluated interview quality, such that high quality applicants had an average accent rating of 2.64,

medium quality applicants had an average accent rating of 2.88, and low quality applicants had an average accent rating of 3.92. Given this relationship between the interview candidates' accents and the apparent quality of their responses, I decided not to use the original video materials. Instead, I decided to identify a subset of applicants and obtain actors that would read their actual responses in hopes of controlling for any variance due to accent.

Of the fifteen interview candidates initially selected, six candidates were further selected on the basis of complete consistency in response quality (2 all high, 2 all medium, 2 all low). One goal in selecting the final six candidates for inclusion in the study was to control for confounding factors, such as race. Thus, the initial plan was to present interviews from six white candidates, and manipulate the race of the candidates in the experiment using photographs. Moreover, as previously discussed, I strove to select interview candidates who were rated consistently high, medium, or low across all portions of the interview. Due to the variance in most white medium-scoring candidates ratings, only five white candidates were selected (i.e., two consistently low-scoring candidates, two consistently high-scoring candidates, and one consistently medium-scoring white candidate). A black candidate's interview was selected for the final medium-scoring interview to bring the entire number of interview stimuli to the original six stimuli.

Since only one presented script was obtained from a black candidate, there is a possibility that ratings of the medium scoring candidates might be confounded by the race of the applicant. Thus, I assessed the extent to which there were systematic differences between ratings of the medium-scoring candidates across participants due to

the applicant race using the written transcript of the applicant's responses. To assess this concern, I conducted a Random Coefficient Model (RCM) in R on ratings of the medium-scoring candidates provided by participants in Study 2. Indeed, there were differences in candidate ratings between these scripts, such that the script obtained from a black candidate was consistently scored as better than the script obtained from a white candidate ( $t(235) = -3.80, p < 0.05$ )<sup>i</sup>. Since the black candidate's rating from the archival study, as evaluated by trained firefighters, was lower than the white candidate's rating in the archival sample, this could indicate that either: a) bias is present in the field ratings, or b) there are quantitative differences in content between the script taken from a black candidate and the script taken from the white candidate.

Six photos (three white, three black) were selected for use in the current study. Online criminal databases were searched using racial criteria (white/black) and age criteria (23-30) to find photographs of seven black men and seven white men. These photographs were then pilot-tested to identify three black and three white men who were equally attractive and masculine, and whose ages were estimated on average to be in the desired range for the study (23-30). Three photos of white men and three photos of black men were selected which were perceived to be similarly attractive and similarly masculine. To affirm that these six photos were perceived to be similarly attractive and masculine, one-way ANOVAs were conducted to assess differences across all six photos. The one-way ANOVAs reveal that there are no differences across all six photos in terms of either attractiveness ( $F(5,108) = 1.45, p > 0.05$ ) or masculinity ( $F(5,108) = 0.73, p > 0.05$ ). Information regarding mean attractiveness, masculinity, and age of the portrayed candidates is depicted in Table 2.

I next assessed differences in ratings that may be confounded by the perceived masculinity and the attractiveness of the photographs employed in the current study. In terms of the influence of attractiveness and masculinity on ratings, more attractive candidates tended to receive higher ratings ( $b = 0.29$ ,  $t(1178) = 3.87$ ,  $p < 0.05$ ), whereas more masculine candidates tended to receive lower ratings ( $b = -0.72$ ,  $t(1178) = -7.08$ ,  $p < 0.05$ ).

The six selected interviews were then transcribed. Of the five original questions, three were selected for inclusion in the current study. These three questions addressed: a) how the applicant would respond in a situation where another firefighter was not pitching in to do his or her fair share, b) how the applicant would react in a group work situation where his or her colleagues were struggling with their assigned tasks, and c) how the applicant would respond to a civilian interruption at the firehouse at 2 a.m.

Six white voice actors (four research assistants, one theater major, and one post-doc) were hired to read each candidate's response to each question. White voice actors were selected in order to avoid possible confounds in language. Actors were provided with coaching on sounding natural when recording each response. Each actor was allowed to have as many takes as necessary to get through each interview response fully. These recordings were then digitally manipulated to decrease the pitch of the actors' voices to make the voices racially ambiguous.

A pilot study was conducted to assess whether or not the pitch manipulation worked. In this pilot study, participants rated each voice in terms of how believable it is that the voice belongs to a white or black male. Specifically, participants were asked: "If you were told that the candidate featured in the recording above was white/black, to what

extent would that be believable to you?” Participants responded to such items on a scale from 1-5, with 1 being “to no extent” and 5 being “to a great extent”.

Only one of the voices was perceived as racially ambiguous ( $t(14) = 0.34, p > 0.05$ , Mean white = 3.47, Mean black = 3.33). All other voices were perceived as being more likely to belong to a white person rather than a black person. Table 3 displays the results of these analyses. Since the voices were not universally perceived as racially ambiguous in the pilot study, I tested whether apparent race of voice interacted with presented race of candidate in the main study using RCM in R.

Candidate vocal profiles were related to perceived quality. In particular, the more “white” a candidate sounds, the higher he was rated ( $b = 0.27, t(1178) = 3.46, p < 0.05$ ), and the more “black” a candidate sounds, the lower he was rated ( $b = -0.11, t(1178) = -2.38, p < 0.05$ ). Moreover, in each case, the vocal profile interacts with the manipulated candidate race in predicting ratings. In particular, black candidates whose voices seem more “white” have the highest ratings ( $b = -0.22, t(1176) = -2.85$ ), and black candidates whose voices seem more “black” have the lowest ratings ( $b = 0.10, t(1176) = 2.13$ ).

In conclusion, aspects of the study design (such as vocal profiles, masculinity, and attractiveness) did impact candidate ratings<sup>ii</sup>. Consequently, it was important to counterbalance these concerns within and across candidates in order to mitigate the impact of these confounds. Next, I describe how the study was designed, and explain how study design was employed to address these confounds.

## **Design**

The current study employed a 3x2x2 between-participant design (3 levels of quality x 2 races of candidate x 2 scripts for each level of quality). That is, the race of the

candidate and quality of the interview first presented to the participant might influence that participant's ratings in a meaningful way due to a potential information order bias (Perrin, Barnett, Walrath, & Grossman, 2001). Thus, interview quality and race were counterbalanced such that each interview (of six) and each race (of two) was presented first exactly once. Thus, in total, twelve conditions (3 levels of quality x 2 races x 2 scripts) were employed in the current study. For a full description of the study design, see Table 4.

Since each condition presents candidates in a different order, the impact of candidate presentation order was assessed in R. Unfortunately, order of presentation was found to affect applicant ratings with the first three candidates rated higher than the last three ( $b = 0.07$ ,  $t(1174) = 3.20$ ,  $p < 0.05$ ) and the third candidate rated somewhat higher than the first two ( $b = -0.11$ ,  $t(1174) = -2.52$ ,  $p < 0.05$ ). However, since presentation of candidates was counterbalanced across conditions, such that each condition had a different script order for the candidates, this set of variables is likely to reduce the power of detecting effects of interest rather than truly confounding the main analyses. In other words, order is not confounded with candidate script, and thus, it is not confounded with candidate quality. Summaries of analyses of main effects of the potential design-related confounds are in Table 5, and summaries of analyses of the interactive effects of these potential confounds with candidate race are provided in Table 6.

In order to account for the design-related confounds discussed above, the stimuli were constructed in the following manner. First, each participant heard only one of the actors for all of the six vocal profiles during the study. In other words, accent differences were held constant within individuals. Moreover, black and white candidates were

matched, as previously discussed, on attractiveness and masculinity, thus minimizing the confounding effect of these factors on candidate race. Lastly, as previously discussed, counterbalancing the order of race and quality of the candidates across conditions helps to neutralize the potential biasing effects of order on discrimination in rating.

### **Procedure**

Student participants were provided introductory information about the study. Specifically, participants were introduced to the interview scenarios and the rating scales, and were informed about the benchmarks used to anchor the rating scales. Participants were asked to be as objective as possible when rating candidates. The script employed to orient the participants to the study is included in Appendix B. After receiving this introductory information, participants then began the online study.

The first screen detailed information on each of the three interview questions and the rating scale benchmarks participants would be using to rate the job candidates. The actual information provided to participants is included in Appendix C. No other training was provided.

Once participants were comfortable with this information, they proceeded to the next screen to begin rating the first job candidate. Each successive page displayed the text of one question, an audio file for one candidate's response to that question, a picture of the candidate, some basic demographic information on the candidate (i.e. race, gender, age), a place to take notes, the rating scale, and the benchmarks to be used for that particular question. A screenshot of one of these pages is in Figure 3. After rating the candidate on a given question, each participant moved forward to the next page, which would display the same candidate and his response to the next question. Thus, each

candidate's responses spanned three pages, with one page per question. Following the pages employed for rating candidates, participants ranked the candidates from 1-6, with 1 being the best and 6 being the worst. Participants then provided their perceptions of the selection process as a whole.

After rating and ranking the candidates, participants supplied information on their personal attitudes towards whites and blacks (Appendix D), measures of a motivation to hide prejudice against whites and blacks (Appendix E), and then filled out demographic information (Appendix F). Participants were debriefed in person.

## **Materials**

**Prejudice toward blacks and prejudice toward whites.** Scales to assess prejudice against blacks and whites were constructed using items with referents that could be meaningfully changed from three different scales: the attitudes towards blacks scale (Brigham, 1993), the modern racism scale (McConahay, 1986), and the updated symbolic racism scale (Henry & Sears, 2002). Both forms of these scales are in Appendix D.

These scales were assessed for construct equivalence. Since the scales employed in the current study were developed from multiple sources, I took a multi-stage approach to assessing construct equivalence. First, I conducted a maximum likelihood exploratory factor analyses to find the overall model. Then, I conducted a series of confirmatory factor analyses to assess the extent to which the specified model fit for both black and white referents.

The scree plots and eigenvalues provided by the exploratory factor analysis indicated that three (for black-referent items) to four factors (for white-referent items)



might be present for each form of the scale. These plots are provided in Figures 4.1 and 4.2. However, orthogonal rotation and inspection of factor loadings indicated that reverse-coded items loaded on separate factors than items in the positive direction. Indeed, prior research reveals that participants simply do not respond the same way to reverse-coded as non-reverse-coded items, leading to spurious factors that are methods-based rather than construct-based (Spector, Van Katwyk, Brannick, & Chen, 1997). Thus, I decided to drop the 5 reverse-coded items from each scale and assess the factor structure of the remaining items for black and white referents. Maximum-likelihood exploratory factor analyses were again conducted on these 7 items. For both black and white referents, scree plots (Figures 5.1 and 5.2) and eigenvalues indicated that the seven items loaded on two factors. The rotated factor solution for whites as a referent is provided in Table 7. The rotated factor solution for blacks as a referent is provided in Table 8. Within both of these analyses, items 1, 4, 5, and 6 (from the symbolic racism and the modern racism scales) loaded on the first factor, and items 8, 10, and 12 (from the attitudes toward blacks scale) loaded on the second factor.

Next, I conducted a series of multi-group factor analyses in MPLUS to assess construct equivalence between black- and white-referent items. First, I estimated an unconstrained model with two factors specified for both black- and white-referent items. Then, I constrained factor loadings one at a time, until the  $\chi^2$  difference between models became statistically significant. Four out of seven possible loadings were able to be constrained before the  $\chi^2$  difference between models became statistically significant. Thus, full measurement equivalence could not be attained. However, when all loadings were constrained to be equal across white- and black-referent items, the model still fit

well ( $\chi^2(31) = 61.90$ ,  $p < 0.05$ ; CFI = 0.94; RMSEA = 0.07; SRMR = 0.07). Thus, while full measurement equivalence could not be attained, these scales have reasonably similar measurement properties, and thus can be considered to be configuration-equivalent, and, to a lesser extent, somewhat metric-equivalent (VandenBerg & Lance, 2000). A summary of the fit statistics for the estimated models is available in Table 9.

Finally, I constructed scales for prejudice against whites and blacks. I first averaged together items within each factor to create a mean for each individual on each of the two factors, for both black- and white-referent items. Then, since the correlations between the two factors were reasonably high (0.52 for white-referent items, 0.73 for black-referent items), I next transformed these factor means into z-scores. I then summed together these z-scores to create an overall prejudice score for each individual on both the white- and black-referent items. I then employed equations specifying the means and variances of linear composites in order to put both of these scales back into their original measurement (Nunnally & Bernstein, 1994). I first divided the summed z-score scales of prejudice by their respective standard deviations. Next, I found the variance of the linear composite using the following equation:

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{ij} \quad \text{Equation 4.1}$$

According to this equation, the variance of the composite is equal to the sum of the variances of the scales in the composite added to two times the covariance of the scales. From this variance, I determined the standard deviation of the linear composite. Next, I multiplied the summed z-score scales by this standard deviation. Finally, I added to these scores the mean of each composite. The means of the composites were found using the following equation:

$$\bar{y} = \bar{y}_1 + \bar{y}_2 \quad \text{Equation 4.2}$$

Thus, the mean of each composite was the sum of the means of the scales in that composite. Finally, since each composite consisted of two scale scores, I obtained the average prejudice measures for each individual by dividing the scores by two. Reliability estimates employing the seven items for each prejudice scale were acceptable (Cronbach's alpha = 0.70 and 0.69, respectively).

**Motivation to hide prejudice.** The external motivation to hide prejudice scale (Plant & Devine, 1998) was employed in the current study as a control variable in order to assess the extent to which individuals may be motivated to hide their prejudicial beliefs. Specifically, prior research on a process called "flexible corrections" indicates that individuals may anticipate their own prejudices and try to correct them in an attempt to hide these prejudices (Petty & Wegener, 1993; Wegener & Petty, 1995).

Two outcomes may result from these attempts. First, individuals may appear to favor those whom they are prejudiced against if they overestimate the extent of their own prejudice, or may appear to penalize the group about whom they do not hold prejudicial beliefs. Thus, I included motivation to hide prejudice in the current study in order to account for these potential concerns. In addition to the traditional motivation to hide prejudice against black's scale, this scale's referents were altered to create a white-centric motivation to hide prejudice scale. Both forms of these scales are available in Appendix E.

To assess the construct equivalence of this scale, a series of confirmatory factor analyses were conducted in MPLUS. First, separate factor analyses on black-referent items and white-referent items were conducted to assess the extent to which a one-factor

structure fits the data. The model specifying one factor fit reasonably well for both white ( $\chi^2(2) = 164.67$ ,  $p < 0.05$ ; CFI = 0.96; RMSEA = 0.21 (CI: 0.19-0.24); SRMR = 0.04) and black ( $\chi^2(2) = 140.36$ ,  $p < 0.05$ ; CFI = 0.95; RMSEA = 0.20 (CI: 0.17-0.23); SRMR = 0.05) referent items separately.

Next, I conducted a series of multi-group factor analyses to assess construct equivalence between black- and white-referent items. First, I estimated an unconstrained model. Then, I constrained factor loadings one at a time, until the  $\chi^2$  difference between models became statistically significant. Two out of four possible loadings were able to be constrained before the  $\chi^2$  difference between models became statistically significant. Thus, full measurement equivalence could not be attained. However, when all loadings were constrained to be equal across white- and black-referent items, the model still fit well ( $\chi^2(7) = 326.33$ ,  $p < 0.05$ ; CFI = 0.95; RMSEA = 0.16; SRMR = 0.05). Thus, while full measurement equivalence could not be attained, these scales have reasonably similar measurement properties, and can be considered configuration-equivalent (same factor structure across referents), and, to a lesser extent, somewhat metric equivalent (same loadings across referents; Vandenberg & Lance, 2000). A summary of the fit statistics for the estimated models is available in Table 10.

Finally, I constructed scales for motivation to hide prejudice against whites and blacks by averaging the four items together within each referent. Reliability estimates for both the overall scale for motivation to hide prejudice with both black- and white-referent items were acceptable (Cronbach's alpha = 0.81 and 0.86, respectively).

## Study 2 Results

### Manipulation Check

As discussed previously, 2 low-quality, 2 moderate-quality, and 2 high-quality candidates were included in the current study. To test whether my manipulation worked, I conducted a RCM to assess the extent to which candidate quality was related to candidate ratings. Indeed, consistent with expectations, the manipulated candidate quality was significantly positively related to candidate ratings ( $b = 0.32$ ,  $t(1178) = 20.77$ ,  $p < 0.05$ ). Thus, the candidate quality manipulation was successful.

### Confounds

I assessed the extent to which demographic variables influenced overall ratings and the relationship between candidate race and ratings. Specifically, I analyzed the impact of age, gender, major (psychology versus non-psychology), religion, political orientation (liberal versus not-liberal), year in college, socio-economic status, and GPA. None of these demographic variables had a significant main effect on ratings, nor did any influence the relationship between candidate race and ratings. Summaries of the main effects analyses are available in Table 11, and summaries of the interaction analyses are in Table 12.

### Tests of Hypotheses

Correlations between participant race, average ratings of black and white candidates, average difficulty parameters for black and white candidates, prejudice against blacks and whites, and motivation to hide prejudice against blacks and whites are displayed in Table 13. As in Study 1, a series of PCM IRT models were fit to the student data for preliminary tests of the hypotheses. Again, I conducted a confirmatory factor

analysis in MPLUS to see if the three ratings tap into an overall “effectiveness” construct. This analysis shows that indeed, all three ratings load on to one factor representing overall effectiveness at or above a standardized loading of 0.5. Given that this model is just-identified, factor loadings are the only way to evaluate the appropriateness of this model, as fit indices cannot be calculated for just-identified models. Since these items are all tapping into the same latent construct of overall effectiveness, I do not investigate individual item differences in the following analyses. Instead, I estimate step difficulties across all five items.

Next, I analyzed responses using IRT to assess differences in step difficulties across raters. Specifically, data was fit to a series of partial credit models (PCMs) using ConQUEST. For the first hypothesis, I performed a series of PCM analyses similar to those conducted in Study 1. Specifically, I first imposed a PCM model that predicts item difficulty as a function of differences between raters. This model assumes that applicant race does not affect rater evaluations of the applicants, but that raters differ in difficulty between Likert scale markers. Indeed, this analysis reveals that rater characteristics clearly impact step difficulty parameters on average ( $\chi^2(211) = 681.59, p < 0.05$ ).

I next imposed a second PCM model on the data. This second model included applicant race as an additive factor. This model assumes that there is an overall effect for applicant race that is consistent across all raters. According to this analysis, rater ( $\chi^2(211) = 8220.43, p < 0.05$ ) predicts difficulty parameters on average. Race of applicant ( $\chi^2(1) = 0.01, p < 0.05$ ), however, does not appear to predict average difficulty parameters.

The third model adds a rater by race interaction to reflect the possibility that raters vary on their usage of the rating scale, given applicant race. The difference between this

model and the two aforementioned models is that it assumes that raters are differentially sensitive to applicant race. As hypothesized, there is a significant rater by applicant race interaction in step difficulty parameters ( $\chi^2(211) = 378.01, p < 0.05$ ). As such, it appears that applicant race and rater characteristics interact to predict step difficulties.

According to Hypothesis 1, one relevant rater characteristic that should predict DTF-for is the rater's race. In other words, raters should be more lenient (easier step difficulty parameters) when rating same-race relative to different-race applicants. Thus, to analyze this hypothesis, I ran another PCM in ConQUEST specifying that rater race interacts with applicant race to predict step difficulty parameters. This equation builds on the prior equations in that it explicitly specifies the relevant characteristics of the rater (race) as interacting with applicant race in predicting step difficulty parameters. Evidence for Hypothesis 1 would be provided by a significant rater by race interaction, provided that the direction of the interaction is consistent with the aforementioned hypothesis.

There was no significant interaction between applicant race and rater race ( $\chi^2(1) = 0.61, p < 0.05$ ) in the prediction of difficulty parameters. However, as previously discussed, the omnibus test does not provide a nuanced view of DTF across the Likert scale markers. Thus, I investigated pair-wise comparisons between rater race and ratee race for each set of step parameters using the estimated standard errors. Further analysis of the interaction between rater race, applicant race, and steps between items reveals interesting differences. Table 14 and Figure 6 depict how black and white raters differentially respond to black and white candidates.

Analysis of black raters' assessment of white and black applicants reveals some statistically significant differences. Specifically, on the first step, black raters were more

lenient toward black applicants relative to white applicants (Step 1:  $t(66) = 3.70, p < 0.01$ ) but more severe toward black applicants relative to white applicants on the second step ( $t(66) = -6.67, p < 0.01$ ). White raters were more lenient toward white applicants relative to black applicants on the first step ( $t(166) = 5.06, p < 0.01$ ). However, there were no significant differences for ratings of white and black applicants by white raters on the second step ( $t(166) = -0.50, p > 0.10$ ).

Overall, there was evidence for differential rater use of the scale as a function of applicant race. While support for Hypothesis 1 was not consistent across the first two steps, there is support for the hypothesis at the first step on the scale. In other words, raters have an easier time moving from a “1” to a “2” when assessing candidates of their own race.

I tested Hypotheses 2 and 3 using LCMM analyses in MPLUS. First, I generated step difficulty parameters for each rater, collapsing across applicant race. In other words, each rater had a set of step difficulty parameters for white applicants and another set of step difficulty parameters for black applicants. These difficulty parameters were then used as dependent variables in the mixture model analyses. Specifically, three effects-coded variables were generated contrasting step difficulty parameters within the Likert scale

Step difficulty parameters for both black and white applicants were nested within raters. Thus, applicant race, the three effects-coded variables, and the interactions between applicant race and the effects-coded variables were specified as within-individual variables, whereas prejudice and rater race were specified as between-individual variables. On the within level, I specified an equation wherein which step



difficulty parameters were predicted by applicant race, the three effects-coded variables representing the location of the difficulty parameter on the Likert scale, and the three interaction terms. Specifically, I estimated the following model on the within level:

$$\delta_{ih} = w_1 \text{apprace} + w_2 \text{contrast1} + w_3 \text{contrast2} + w_4 \text{contrast3} + w_5 \text{apprace} * \text{contrast1} + w_6 \text{apprace} * \text{contrast2} + w_7 \text{apprace} * \text{contrast3} \quad \text{Equation 4.1}$$

The within-level equation was initially specified to vary across latent classes. I estimated these equations for models with 2 to 6 latent classes, without including predictors. I then compared the log-likelihood criteria, Bayesian Information Criteria (BIC), and Akaike Information Criteria (AIC) across these five models to determine which model best fits the data. Specifically, the “best” model would be one in which the BIC and AIC were minimized. Additionally, the product of two and the difference between two models’ likelihood criteria follows a chi-square distribution with degrees of freedom being the difference in degrees of freedom between the two models. Using the following three methods to assess the fit statistics, it is apparent that five (LL = -2298.28, AIC = 4686.56, BIC = 4918.231) latent classes fit the data better than four latent classes (LL = -2351.84, AIC = 4775.67, BIC = 4961.01). Additionally, estimating six latent classes (LL = -2298.28, AIC = 4704.56, BIC = 4982.57) does not significantly improve fit over five estimated latent classes. Thus, it appears that five latent classes do fit the data best.

Next, I then predicted class membership for a model specifying five latent classes using prejudice against blacks, prejudice against whites, rater race, and motivation to hide prejudice against blacks and whites. Specifically, I predicted class membership using the following equation at the between level:

$$c = v1race + v2whiteprejudice + v3blackprejudice + v4mothideblackprej + v5mothidewhiteprej \quad \text{Equation 4.2}$$

The results of this analysis show that each of the five classes constitutes between 5% and 65% of the overall sample. That is, the third class constitutes 65% of the overall sample, whereas the fourth class constitutes 11% of the sample, and fifth class constitutes 9% of the overall sample. The first class represents 10%, and the second class represents 5% of the overall sample. Tables 15 – 19 provide summaries of the regression coefficients that correspond to the relationship between each of the within-level predictors and the difficulty parameters within the five classes.

In all five latent classes, at least one interaction between candidate race and the effects-coded scale location variables was significant, rendering interpretations of main effects in the context of the higher-order interactions meaningless. The patterns of these interactions vary widely across classes. Using the estimated intercept and regression parameters, estimated marginal means were computed for white and black difficulty parameters for each step within each class. Figures 7.1-7.5 depict these estimated marginal means. Since at least one interaction was significant within each latent class, it is challenging to directly interpret the estimated parameters. Thus, the depictions of the estimated marginal means were employed in conjunction with the statistical results to assess the underlying nature of DTF present in each group. Hypothesis 2 will be supported if five distinct patterns emerge in the data: a) no DTF, b) DTF-against blacks and not for whites, c) DTF-against blacks and DTF-for whites, d) DTF-for whites and not against blacks, e) DTF-against whites and not toward blacks.

In the first latent class, raters are much stricter in their ratings of black applicants when moving from a marker of “1” to a marker of “2”. Conversely, raters in this class use

the scale relatively consistently for white candidates. Given the visual inspection, and two significant interactions terms, it appears that individuals in this latent class exhibit DTF against black candidates, but not in favor of white candidates.

In the second latent class, raters are stricter in their ratings of white applicants when moving from a marker of “1” to a marker of “2” and from a marker of “2” to a marker of “3”. However, these raters are also strict toward black candidates when moving from a marker of “1” to a marker of “2”. While these individuals are universally harsher on both black and white candidates, there does seem to be a noticeable difference in how black and white candidates are evaluated in terms of usage of the 2<sup>nd</sup> and 3<sup>rd</sup> Likert scale markers. Specifically, these participants are harsher on white candidates at this juncture. Thus, this latent class seems to be comprised of individuals who exhibit DTF against white applicants, but also DTF somewhat in favor of black applicants.

In the third latent class, only one of the three interaction terms is significant. Moreover, raters in this class do not seem to substantively employ the scale markers differentially for white versus black candidates. As such, this class of individuals appears to have no DTF. It is important to note that this is the only class wherein which the effects-coded variable for the first step is substantively below zero. That is, this is the class wherein which raters seemed to avoid giving applicants a rating of “1”.

In the fourth latent class, raters are much stricter in their ratings of black applicants when moving from a marker of “2” to a marker of “3”. Conversely, raters in this class appear to employ the markers of the scale consistently for white applicants. Consequently, it appears that individuals in this latent class exhibit DTF-against black candidates, but no DTF toward white candidates.

In the fifth latent class, raters are much stricter in their ratings of white applicants when moving from a marker of “1” to a marker of “2”. Conversely, these raters seem to employ the scale very consistently when rating black candidates. As such, this latent class seems to be comprised of individuals who exhibit DTF-against white applicants, but not in favor of black applicants.

In sum, five latent classes were uncovered. These classes were comprised of individuals who exhibited varied patterns of DTF: a) DTF against whites and DTF for blacks (class 2), b) DTF against whites but not in favor of blacks (class 5), c) DTF against blacks but not in favor of whites (classes 1 and 4), and d) no DTF (class 3). Of these five classes, three were predicted in Hypothesis 2. Specifically, Hypothesis 2 specified that a class of individuals who exhibit DTF-for whites but not against blacks should be found, and that a class of individuals who exhibit DTF-against blacks and DTF-for whites should be found. Additionally, Hypothesis 2 did not specify that there should be a class comprised of individuals exhibiting DTF-against whites and DTF-for blacks. Overall, there is some support for Hypothesis 2, with the exception of the portions of the hypothesis that predicted DTF-for white candidates. Instead, it seems that DTF-against blacks primarily drives discriminatory responding toward black applicants, rather than DTF in favor of whites.

Hypothesis 3 predicted that rater race and rater prejudice would predict latent class membership. The second class (which appeared to exhibit DTF against whites and somewhat in favor of blacks) is composed of significantly more white individuals than any other class. Additionally, the fifth latent class (which seemed to exhibit DTF against whites, but not in favor of blacks) is composed of the largest number of black individuals,

especially compared to either the first or third classes, although these differences are not statistically significant. These findings run contrary to the hypothesis that in-group membership would predict DTF-for same-race candidates for white participants, but supports the hypothesis that in-group membership would predict DTF-against whites for black participants. Moreover, prejudice does not predict class membership at all. Consequently, there is limited support for Hypothesis 3.

To explore why prejudice might not have predicted class membership, I also assessed the extent to which classes differed in terms of motivation to hide prejudice. Indeed, an analysis of motivation to hide prejudice reveals that motivation to hide prejudice against blacks distinguishes between class 2 and all other classes. Specifically, class 2 is comprised of individuals with a higher motivation to hide prejudice against blacks than any other class. Thus, class two is composed primarily of white individuals who are highly motivated to suppress prejudice against blacks. As discussed previously, the flexible corrections model specifies that individuals' awareness of potential prejudices may prompt them to try to suppress these prejudices (Petty & Wegener, 1993; Wegener & Petty, 1995). Consequently, it is reasonable to conclude that the unusual response patterns in this class could be a result of conscious efforts to suppress prejudice.

### **Study 2 Discussion**

In Study 2, I attempted to replicate and extend the results of Study 1 in a lab context. Specifically, I sought to predict DTF-for using in-group belonging (e.g. a match between applicant race and rater race) and DTF-against using prejudice. The results of the current study revealed evidence for disparate usage of scales based on race of the applicant. Moreover, the current study showed distinct patterns of DTF-against and no

DTF could be detected and predicted using latent class mixture modeling. Interestingly, the results of Study 2 indicate that DTF-against may be driven more by in-group belonging than by prejudice. Additionally, DTF-for does not seem to be operating. Indeed, it may be that a motivation to hide prejudice drives responding for some participants more than reported prejudice. One final finding from Study 2 that is particularly compelling is that there were two classes of individuals who exhibited DTF against black applicants, but not in favor of white applicants. These individuals not only exhibited DTF at different points in the scale, but also possessed different individual characteristics. Hence, DTF-against may not be as simple of a phenomenon as previously estimated.

While the results of this second study are illuminating, it is possible that the classes of individuals found were due to unique variations within the sample, instead of due to true differences in the underlying populations. In particular, it might be that college students exhibit different patterns of prejudice and motivation to hide prejudice than non-college students (Henry, 2008), which may have driven different patterns of responding and class profile.

Additionally, as discussed, DTF-for an in-group was not found in the current study. It may be that intergroup bias in interview contexts is primarily driven by DTF-against, due to perceived competition in the job application process (Hewstone et al., 2002). However, it is also possible that DTF-for was not found due to this study's operationalization of in-group belonging. Specifically, it may be that a match between the rater's race and the applicant's race is not enough to predict DTF-for. That is, a particularly salient individual trait that may lead to intergroup prejudice and DTF-for

one's own group is ethnic identity (Phinney, 1992). Thus, I next sought to replicate and extend the results of Study 2 in an adult sample in order to systematically address these concerns. Specifically, adult participants engaged in the same experiment that student participants completed. Moreover, I collected additional measures of ethnic identity and additional prejudice scales in Study 3 to assist in disentangling some of the drawbacks of the second study. This study is outlined in full next.

### Study 3

In the intergroup bias literature, in-group identification is a critical driver of intergroup bias (Hewstone et al., 2002). In the context of the current study, a relevant form of in-group identification is ethnic identity (Phinney, 1992). Ethnic identity is the extent to which an individual's ethnic group belonging is important to their self-identity (Phinney, 1992). Given that ethnic identity is a salient form of in-group identification for the purposes of the current research, it is possible that DTF-for the in-group in white individuals does not occur except for those who have a strong ethnic identity.

Consequently, I hypothesize the following:

*Hypothesis 4: Ethnic identity, rater race, prejudice, and motivation to hide prejudice will predict patterns of responding (in the form of latent class membership), such that:*

- a. In classes where white raters are predominant and the raters have high ethnic identity, there will be DTF-for white applicants.*
- b. In classes where black raters are predominant and the raters have high ethnic identity, there will be DTF-against white applicants.*
- c. Raters with high levels of prejudice will exhibit DTF-against the race toward which the prejudicial beliefs are held.*

- d. Raters with high levels of motivation to hide prejudice will exhibit patterns of responding where they appear to exhibit DTF against their own race.*

### **Study 3 Method**

#### **Participants**

Participants were 182 adults, recruited through Mechanical Turk. Mechanical Turk is an online community wherein which “requesters” and “workers” can connect. Specifically, requesters post work (or studies) that need to be completed, and provide necessary information about this work. Workers may then select tasks that they wish to complete for a small fee. The current study was posted on Mechanical Turk, and white and black non-student participants were recruited through this means. Participants were awarded \$1.00 for completing the task. Before accepting and completing the task, participants knew that they would receive \$1.00, and also were informed that the task takes 30 minutes to an hour to complete. Preliminary research on the population of Mechanical Turk workers reveals that these workers are primarily young adults, come from a variety of educational backgrounds, are fairly evenly split between unemployed, employed part-time, and employed full-time, and slightly over 50% female (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). In other words, while Mechanical Turk workers are fairly homogenous in some ways (e.g. age), they also exhibit a great deal of diversity (e.g. education, employment, and gender).

This particular sample was 33.5% black ( $n = 61$ ) and 66.5% white ( $n = 121$ ). Additionally, the adult sample was 64.8% female ( $n = 118$ ). The majority of participants completed their undergraduate degree (42.3%,  $n = 77$ ), followed by community college (24.2%,  $n = 44$ ), secondary school (18.7%,  $n = 34$ ), and graduate school (13.7%,  $n = 25$ ).



Of the college majors reported, the largest portion of participants were business majors (12.1%,  $n = 22$ ). No other major was represented by more than 5% of the total sample. Participants' High School GPAs ranged from 2.00 to 4.16 (mean = 3.47, stdev = 0.48), and college GPAs ranged from 2.40 to 4.00 (mean = 3.46, stdev = 0.38). Participant ages ranged from 18 to 63 (mean = 33.38, stdev = 10.05).

The adult sample is 64.3% Christian ( $n = 117$ ), 9.9% agnostic ( $n = 18$ ), 8.8% spiritual but not religious ( $n = 16$ ), and 8.2% Atheist ( $n = 15$ ). No other religious group was represented by at least 5% of the sample. Of the participants, 42.9% were democrats ( $n = 78$ ), 26.9% of the participants were independent ( $n = 49$ ), 17.0% were republican ( $n = 31$ ), and 7.1% were Libertarian ( $n = 13$ ). No other political affiliation was represented by at least 5% of the sample. Adult participants are largely middle-class (74.2%,  $n = 135$ ), followed by lower-class (24.2%,  $n = 44$ ). For the majority of participants, English is their native language (94.0%,  $n = 171$ ), and they were born in the United States (90.7%,  $n = 165$ ).

Adult participants currently reside in 39 out of 51 states (including the District of Columbia). States in which adult participants did not reside are Alaska, Arkansas, Delaware, Hawaii, Idaho, Maine, New Hampshire, North Dakota, Rhode Island, South Dakota, Vermont, and Wyoming. In terms of the states represented in this sample, 7.1% of participants currently live in Florida ( $n = 13$ ), 6.0% of participants live in each of Georgia and Illinois ( $n = 11$ ), and 5.5% of participants live in each of Michigan and New York ( $n = 10$ ). No other state was represented by at least 5% of the sample.

In terms of current employment, 72.5% of participants work full-time ( $n = 132$ ). Participants largely come from the healthcare industry (11.5%,  $n = 21$ ), education,

training, and library work (11.0%,  $n = 20$ ), business and financial operations (7.1%,  $n = 13$ ), and computer/mathematical work (5.5%,  $n = 10$ ). Additionally, 6.6% of participants self-identify as homemakers ( $n = 12$ ). No other industry was represented by at least 5% of the sample. Hours worked per week ranged from 8 to 80 (mean = 38.33, stdev = 12.09).

Participants largely did not have experience working in firefighting, EMT, or paramedic industries. However, 22.5% of adult participants ( $n = 41$ ) had some experience interviewing job candidates in the past. Consequently, I will assess the extent to which prior interview experience influences ratings of candidates.

### **Design**

The same six interviews, three questions, and overall design employed in Study 2 was again utilized in the current study. Specifically, twelve conditions (3 levels of quality x 2 races x 2 scripts) were employed in the current study. For a full description of the study design, see Table 2.

### **Procedure**

Adult participants were sent to a website which displayed a consent form. In order to proceed with the study, adult participants had to provide their consent at the bottom of this form. The remainder of the study proceeded as in Study 2, with the exception of the debriefing information, which was also provided online. That is, participants were introduced to the scenarios, and the purpose of the provided benchmarks was reviewed. Participants were asked to be as objective as possible when rating candidates. Then, participants proceeded to listen to and rate six candidates for an entry-level firefighting position. After rating each candidate individually, participants ranked the candidates from 1-6, with 1 being the best and 6 being the worst. They then provided their perceptions of

the selection process as a whole. Finally, participants supplied information on their personal attitudes towards whites and blacks (Appendix D and Appendix G), measures of a motivation to hide prejudice against whites and blacks (Appendix E), an ethnic identity measure (Appendix H), and then filled out demographic information (Appendix I).

Participants were provided an online debriefing form.

## **Measures**

**Prejudice toward blacks and prejudice toward whites.** The same scales employed in Study 2 were utilized again in Study 3 to assess prejudice toward blacks and whites. Again, these scales were constructed using items with referents that could be meaningfully changed from three different scales: the attitudes towards blacks scale (Brigham, 1993), the modern racism scale (McConahay, 1986), and the updated symbolic racism scale (Henry & Sears, 2002).

As previously discussed, the prejudice scales employed in Study 2 did not predict latent class category membership. This finding may be due to the overpowering effects of motivation to hide prejudice, or it may be due to construct issues. Specifically, the reliabilities of the prejudice scales for both blacks (0.70) and whites (0.69) were fairly low, and full metric equivalence was not achieved. Consequently, I included a second set of prejudice scales in the current study. Specifically, while the scales employed in Study 2 were developed from prejudice scales against blacks, I added scales developed from a prejudice measure against whites. That is, I included an additional 15 items from Johnson and Lecci's (2003) white prejudice scale whose referents could be meaningfully changed. White and black-referent forms of both of these scales are in Appendix G.

First, to replicate results from Study 2, I analyzed the data for items from the black and white prejudice scales employed in Study 2. Specifically, I assessed the factor structure of items 1, 4, 5, 6 (from the symbolic racism and modern racism scale), and 8, 10, and 12 (from the attitudes toward blacks scale) from the first 12 prejudice items using Confirmatory Factor Analysis (CFA) in MPLUS. As in Study 2, I first assessed the extent to which a two factor structure best fits the data. The model specifying two factors fit best for white ( $\chi^2(13) = 38.19$ ,  $p < 0.05$ ; CFI = 0.91; RMSEA = 0.10; SRMR = 0.06) referent items. However, the model specifying three factors fit best for black ( $\chi^2(11) = 28.93$ ,  $p < 0.05$ ; CFI = 0.95; RMSEA = 0.10; SRMR = 0.05) referent items. Results from these analyses are depicted in Table 20. The inconsistencies in findings between Study 2 and Study 3 in terms of the structure of this scale may indicate that the scales' properties are not stable. Despite the instability in results, I will move forward with analyses of the 2-factor structure in order to allow for comparisons across Study 2 and Study 3 in terms of the predictive validity of this particular prejudice measure.

Next, I conducted a series of multi-group factor analyses to assess construct equivalence between black- and white-referent items as in Study 2. First, I estimated an unconstrained model. Then, I constrained factor loadings one at a time, until the  $\chi^2$  difference between models became statistically significant. Five out of seven possible loadings were able to be restrained before the  $\chi^2$  difference between models became statistically significant. Thus, full measurement equivalence, as found previously, could not be attained. However, when all loadings were constrained to be equal across white- and black-referent items, the model still fit reasonably well ( $\chi^2(31) = 90.57$ ,  $p < 0.05$ ; CFI = 0.90; RMSEA = 0.10; SRMR = 0.07). Thus, while full measurement equivalence could

not be attained, the analysis of these scales reveals fairly similar measurement properties, and can be seen as configuration-equivalent (Vandenberg & Lance, 2000). Moreover, the properties of these scales in terms of ability to constrain factor loadings are comparable to those found in the student sample. A summary of the fit statistics for the estimated models is available in Table 21.

My final step for generating the scales previously employed in Study 2 was to create the prejudice scales derived from these seven items. Again, I first averaged together items within each factor to create a mean for each individual on each of the two factors, for both black- and white-referent items. Then, since the correlations between the two factors were high (0.55 for white-referent items, 0.84 for black-referent items), I next transformed these factor means into z-scores. I then summed together these z-scores to create an overall prejudice score for each individual on both the white- and black-referent items. As in Study 2, I then employed equations specifying the means and variances of linear composites in order to put both of these scales back into their original measurement. I first divided the summed z-score scales of prejudice by their respective standard deviations. Next, I multiplied these scales by the standard deviation of the linear composite. Then, I added to these scores the mean of each composite. Finally, since each composite consisted of two scale scores, I obtained the average prejudice measures for each individual by dividing the scores by two. Reliability estimates across the seven items for prejudice against blacks and prejudice against whites were acceptable, and comparable to those obtained in the student sample (Cronbach's alpha = 0.79 and 0.71, respectively).

Next, I sought to find a metric-equivalent scale employing the additional prejudice items included in the current study. I again conducted this analysis in two general steps. First, I conducted maximum likelihood exploratory factor analyses on all items in order to determine the overall model. Then, I conducted CFAs in Mplus to assess the extent to which the overall model fit both black- and white-referent items.

Analysis of the scree plot from the maximum likelihood exploratory factor analyses (Figures 8.1 and 8.2) provided support for the idea that there was one factor underlying the data for both referents. Thus, a one factor solution was forced for both black and white-referent items. The results of these analyses are provided in Table 22. In order to ensure that the included items truly assessed one underlying prejudice measure, prejudice items were dropped if they did not load very highly on either the white- or black-referent factor. Specifically, items from the second set of prejudice scales were dropped if they loaded onto either the white- or black-referent factor at 0.45 or below. Thus, items 2 and 8 from this scale were eliminated from further analysis.

Next, I conducted CFAs in MPLUS on the remaining 13 items. A one-factor CFA did not fit the data particularly impressively for either black- ( $\chi^2(90) = 371.30^*$ ,  $p < 0.05$ ; CFI = 0.83; RMSEA = 0.13; SRMR = 0.07) or white-referent ( $\chi^2(90) = 574.43^*$ ,  $p < 0.05$ ; CFI = 0.74; RMSEA = 0.17; SRMR = 0.10) items. Literature on factor analyses indicates that fit indices can sometimes suffer when items do not display multivariate normality, especially in cases where there are larger numbers of indicators for each factor (Hau & Marsh, 2004). Consequently, researchers encourage the use of “parcels”—combinations of items within the scale (West, Finch, & Curran, 1995). I thus created five parcels of items—three contained three items each, and two of these parcels contained

averages of two items each. I then conducted CFAs using the item parcels. In this case, a one-factor CFA fit the data extremely well for both black- ( $\chi^2(5) = 12.41^*$ ,  $p < 0.05$ ; CFI = 0.99; RMSEA = 0.09; SRMR = 0.02) and white-referent ( $\chi^2(5) = 2.05$ ,  $p > 0.05$ ; CFI = 1.00; RMSEA = 0.00; SRMR = 0.01) items. The results of these analyses are depicted in Table 23. Consequently, it is possible that the prior poor model fit was a reflection of the violation of the assumption of multivariate normality rather than truly poor model fit.

After affirming a one-factor solution fit well for white- and black-referent models, I conducted a series of multi-group CFAs to assess measurement equivalence using the item parcels. To this end, I estimated an unconstrained model. Then, I constrained factor loadings one at a time, until the  $\chi^2$  difference between models became statistically significant. Three out of five possible loadings were able to be constrained before the  $\chi^2$  difference between models became statistically significant. The items in the parcels that could not be constrained (items 3, 4, 7, 13, and 14) were dropped. A summary of these analyses is provided in Table 24.

The difference between the updated three-parcel model with unconstrained loadings and the three-parcel model with fully constrained loadings is non-significant ( $\chi^2(2) = 3.05$ ,  $p > 0.05$ ). However, when the intercepts are constrained to be equal, the difference between models is significant ( $\chi^2(2) = 16.12$ ,  $p < 0.05$ ), indicating that while these scales show metric equivalence (Vandenberg & Lance, 2000), full construct equivalence is not obtained. Finally, I averaged together scores on the remaining items to create overall averages for the white and black prejudice scales. Reliability estimates for both the overall scale for prejudices against blacks and prejudice against whites were quite good (Cronbach's alpha = 0.88 and 0.88, respectively).

**Motivation to hide prejudice.** The external motivation to hide prejudice scale (Plant & Devine, 1998) was again employed in the current study in order to provide control for individuals who may be motivated to hide their prejudicial beliefs. Both forms of these scales are available in Appendix E.

CFAs were again conducted in MPLUS to affirm that the data collected on external motivation to hide prejudice in this study exhibits similar properties to the data collected on external motivation to hide prejudice in Study 2. First, separate factor analyses on black-referent items and white-referent items were conducted to assess the extent to which a one-factor structure fits the data. The model specifying one factor fit well for both white ( $\chi^2(2) = 4.20$ ,  $p > 0.05$ ; CFI = 1.00; RMSEA = 0.08 (CI: 0.00-0.18); SRMR = 0.02) and black ( $\chi^2(2) = 14.65$ ,  $p < 0.05$ ; CFI = 0.95; RMSEA = 0.19 (CI: 0.11-0.28); SRMR = 0.05) referent items separately.

Next, I conducted a series of multi-group factor analyses to assess the construct equivalence between black- and white-referent items. Three out of four possible loadings were able to be constrained before the  $\chi^2$  difference between models became statistically significant. Thus, full measurement equivalence could not be attained. However, when all loadings were constrained to be equal across white- and black-referent items, the model still fit well ( $\chi^2(7) = 28.49$ ,  $p < 0.05$ ; CFI = 0.97; RMSEA = 0.13; SRMR = 0.06). As such, while full measurement equivalence could not be attained, these scales are configuration-equivalent (Vandenberg & Lance, 2000), which is comparable to the findings in Study 2. A summary of the fit statistics for the estimated models is available in Table 25.



Finally, I constructed scales for prejudice against whites and blacks by averaging the four items together within each referent. Reliability estimates for both the overall scale for motivation to hide prejudice with both black- and white-referent items were acceptable (Cronbach's alpha = 0.77 and 0.86, respectively).

**Ethnic identity.** The ethnic identity scale developed by Phinney (1992) was employed in the current study. Since prior work indicated that a one-factor solution adequately described the data, I ran a CFA to affirm the fit of the one-factor solution. As with the prejudice scales, the fit for the one factor solution was not ideal ( $\chi^2(54) = 275.72$ ,  $p < 0.05$ ; CFI = 0.83; RMSEA = 0.15; SRMR = 0.08). Consequently, I again employed item parcels. I created 4 item parcels, with three items in each parcel. When I re-ran the CFA using these parcels, the fit of the one factor solution was excellent ( $\chi^2(2) = 0.16$ ,  $p > 0.05$ ; CFI = 1.00; RMSEA = 0.00; SRMR = 0.00). Table 26 displays the results of this analysis. Consequently, it is possible that the prior poor model fit was a reflection of the violation of the assumption of multivariate normality rather than truly poor model fit.

Next, I conducted a series of multi-group factor analyses to assess the construct equivalence of the ethnic identity scale between black and white participants. All four possible loadings were able to be restrained without the  $\chi^2$  difference between models becoming statistically significant. Moreover, the intercepts of the items, the residual variances of the items, and the variance of the factors were all able to be constrained without the  $\chi^2$  difference between models becoming statistically significant. Consequently, it appears that ethnic identity displays scalar equivalence between black and white participants (Vandenberg & Lance, 2000). A summary of these analyses is provided in Table 27. Last, I created overall scores for ethnic identity by averaging

together the twelve items within the scale. The reliability of the overall ethnic identity scale was excellent (Cronbach's alpha = 0.91).

### **Study 3 Results**

#### **Manipulation Check**

As in Study 2, 2 low-quality, 2 moderate-quality, and 2 high-quality candidates were included in the current study. To test that this manipulation worked, I conducted a RCM to assess the extent to which candidate quality was related to candidate ratings. Indeed, consistent with expectations, candidate quality was significantly positively related to candidate ratings ( $b = 0.24$ ,  $t(919) = 16.27$ ,  $p < 0.05$ ). Thus, the quality manipulation was successful.

#### **Confounds**

I assessed, as in Study 2, the extent to which demographic variables influenced either overall ratings, or the relationship between candidate race and ratings. Specifically, I analyzed the impact of age, gender, high school GPA, college GPA, major (business versus non-business), religion, political orientation (democratic versus not-democratic), socio-economic status, highest level of education, whether or not participants had English as their native language, whether or not participants were born in the United States, full-time work status, hours worked per week, prior experience as an EMT, paramedic, or firefighter, and prior experience interviewing candidates.

Significant main effects on overall ratings were apparent in only one of these analyses. Specifically, individuals with no secondary school education rate candidates lower than those with secondary school education ( $b = -0.46$ ,  $t(179) = -3.17$ ,  $p < 0.05$ ), and both of these groups rate candidates lower than individuals with post-secondary

education ( $b = -0.48$ ,  $t(179) = -2.70$ ,  $p < 0.05$ ). Additionally, socio-economic status interacts with apparent candidate race in predicting ratings, such that individuals who self-identify as “upper class” tend to exacerbate the differences between white and black candidates, with black candidates being rated higher ( $b = 0.26$ ,  $t(917) = 2.98$ ,  $p < 0.05$ ). Summaries of the main effects analyses are available in Table 28, and summaries of the interaction analyses are in Table 29. Since these analyses reveal that demographic variables may influence mean ratings, I will conduct a post-hoc assessment of the demographic make-up of each latent class to assess the extent to which such variables also influence patterns of ratings.

### **Tests of Hypotheses**

Correlations between participant race, average ratings of black and white candidates, average difficulty parameters, ethnic identity, prejudice against blacks and whites, and motivation to hide prejudice against blacks and whites are displayed in Table 30. As in study 1, a series of PCM IRT models were fit to the student data for preliminary tests of the hypotheses. Again, I conducted a confirmatory factor analysis in MPLUS to see if the three ratings tap into an overall “effectiveness” construct. This analysis shows that indeed, all three ratings load on to one factor representing overall effectiveness at or above a standardized loading of 0.5. Given that this model is just-identified, factor loadings are the only way to evaluate the appropriateness of this model, as fit indices cannot be calculated for just-identified models. Since these items are all tapping into the same latent construct of overall effectiveness, I do not investigate individual item differences in the following analyses. Instead, I estimate step difficulties across all three items.

Next, I analyzed responses using IRT to assess differences in step difficulties across raters. Specifically, data was fit to a series of partial credit models (PCMs) using ConQUEST. For the first hypothesis, I performed a series of PCM analyses similar to those conducted in Studies 1 and 2. Specifically, I first imposed a PCM model that predicts item difficulty as a function of differences between raters. This model assumes that applicant race does not affect rater evaluations of the applicants, but that raters differ in difficulty between Likert scale markers. Indeed, this analysis reveals that rater characteristics clearly impact step difficulty parameters on average ( $\chi^2(191) = 464.51, p < 0.05$ ).

I next imposed a second PCM model on the data. This second model included applicant race as an additive factor. This model assumes that there is an overall effect for applicant race that is consistent across all raters. According to this analysis, rater ( $\chi^2(188) = 8191.68, p < 0.05$ ) predicts difficulty parameters on average. Race of applicant ( $\chi^2(1) = 2.69, p < 0.05$ ), however, does not appear to predict average difficulty parameters.

The third model adds a rater by race interaction to reflect the possibility that some raters exhibit different types of DTF than others. The difference between this model and the two aforementioned models is that it assumes that raters are differentially sensitive to applicant race. Indeed, consistent with findings from the prior studies, there is a rater by applicant race interaction in step difficulty parameters ( $\chi^2(188) = 320.63, p < 0.05$ ). As such, it does appear that applicant race and rater characteristics interact to predict step difficulties.

According to Hypothesis 1, one relevant rater characteristic that should predict DTF-for is the rater's race. In other words, raters should be more lenient (easier step

difficulty parameters) when rating same-race relative to different-race applicants. Thus, to analyze this hypothesis, I ran another PCM in ConQUEST specifying that rater race interacts with applicant race to predict step difficulty parameters. This equation builds on the prior equations in that it explicitly specifies the relevant characteristics of the rater (race) as interacting with applicant race in predicting step difficulty parameters. Evidence for Hypothesis 1 would be provided by differential step difficulty parameters between groups, as assessed using standard errors.

There is no significant interaction between applicant race and rater race ( $\chi^2(1) = 2.44, p > 0.05$ ) in predicting difficulty parameters on average. As previously discussed, however, this test does not provide a detailed view of how raters employ different portions of the Likert scale across groups. Indeed, further analysis of the interaction between rater race, applicant race, and steps between items, however, reveals interesting differences. Table 31 and Figure 9 depict how black and white raters differentially respond to black and white candidates.

Analysis of black raters' assessment of white and black applicants reveals some apparent statistically significant differences, based on the estimated standard errors. Specifically, it appears that black raters are more lenient toward black applicants relative to white applicants in the first step ( $t(60) = 3.42, p < 0.01$ ). In other words, it is "easier" for black raters to progress from a rating of "1" to a rating of a "2" when rating black applicants. This finding corresponds with the outcome of the same analysis in Study 2—black raters appear to be more lenient toward black applicants. There were no significant differences for black raters' difficulty parameters on the second step for black and white applicants ( $t(60) = -1.33, p > 0.10$ ).

Conversely, analysis of white raters' assessment of white and black applicants showed that white raters were more lenient toward black applicants compared to white applicants ( $t(120) = -2.83, p < 0.01$ ) on the first step. There were no significant differences for white raters on the second step ( $t(120) = -0.20, p > 0.10$ ). In other words, it is slightly "easier" for white raters to progress from a rating of "1" to a rating of a "2" when rating black applicants. This result is contrary to the finding in Study 2, where white raters were more lenient toward white candidates.

I sought to replicate the latent classes initially found in Study 2 and to test Hypothesis 4 using LCMM analyses in MPLUS. As before, I generated step difficulty parameters for each rater, collapsing across applicant race. These difficulty parameters were then used as dependent variables in the mixture model analyses. Step difficulty parameters for each race were nested within raters. Additionally, three effects-coded variables were generated to represent the four thresholds in the Likert scale. Thus, applicant race, these three effects-coded variables, and three interactions between the effects-coded variables and applicant race were within-individual variables. On the within level, I specified an equation wherein which step difficulty parameters were predicted by applicant race, the three effects-coded variables representing the location of the difficulty parameter on the Likert scale, and the three interaction terms.

The within-level equation was specified, initially, to vary across latent classes. I estimated these equations for models specifying between 2 to 6 latent classes, with no predictors. I then compared the log-likelihood criteria, Bayesian Information Criteria (BIC), and Akaike Information Criteria (AIC) across these five models to determine which model best fits the data. As in Study 2, five ( $LL = -1838.61, AIC = 3767.21, BIC =$

3989.26) latent classes fit the data better than four latent classes (LL = -2068.41, AIC = 4208.81, BIC = 4386.45) or six latent classes (LL = -1861.49, AIC = 3830.99, BIC = 4097.44). Thus, it appears that five latent classes do fit the data best.

Next, I predicted class membership for a model specifying five latent classes using prejudice against blacks, prejudice against whites, race, motivation to hide prejudice against blacks and whites, and ethnic identity. I first conducted this analysis using the prejudice measures employed in Study 2. Then, I conducted this analysis using the construct-equivalent prejudice measures developed in the current study. The patterns of responding within each of the five classes are stable across these analyses. However, the patterns of significance in terms of predictors of the latent classes differ. Thus, when discussing the patterns of responses toward blacks and whites within each class, I will employ the results of the analysis that used the construct-equivalent measures of prejudice. However, when discussing characteristics of each of these five classes, I will discuss both models.

The third latent class represented 66% of the overall sample. This class proportion is almost identical to the largest identified class in the student sample. The second latent class accounted for 13% of the overall sample, the first and third latent classes accounted for 8% of the overall sample, and the fifth class accounted for 6% of the overall sample. All of these percentages are comparable to those found in the student sample. Tables 32 – 36 provide summaries of regression coefficients predicting difficulty parameters within each of the different classes.

Patterns of interactions between candidate race and the effects-coded variables specifying the different markers on the Likert scale vary widely across classes. Using the

calculated intercept and regression parameters, estimated marginal means were computed for white and black difficulty parameters for each step within each class. Figures 10.1-10.5 depict these estimated marginal means. As with the student sample, depictions of the estimated marginal means were employed in conjunction with the statistical results to assess the underlying nature of DTF present in each group. Replication of the latent classes uncovered in Study 2 will be obtained if five distinct patterns emerge in the data: a) no DTF, b) DTF-against blacks and not for whites (at step 1), c) DTF-against blacks and not for whites (at step 2), d) DTF-for blacks and DTF-against whites, and e) DTF-against whites and no DTF toward blacks.

In the first latent class, raters are much stricter in their ratings of black applicants when moving from a marker of “1” to a marker of “2”. While raters in this class are more lenient toward white applicants at this juncture, they are also much stricter when moving from a marker of “2” to a marker of “3” when rating white candidates. It appears that individuals in this latent class exhibit DTF against black candidates, but also somewhat against white candidates. This class bears some similarities in terms of their responses to black applicants to the second class found in the student sample. However, individuals in this class differ from the class found in the student sample in terms of their responses to white candidates. In the student sample, individuals classified as part of the most similar class also seemed to exhibit DTF against whites early on in the scale—this pattern is not identically repeated in the adult sample.

In the second latent class, raters are stricter in their ratings of white applicants when moving from a marker of “1” to a marker of “2”. Conversely, these raters seem to use scale markers equally for black candidates. Thus, this latent class seems to be



comprised of individuals who exhibit DTF against white applicants, but not in favor of black applicants. Consequently, this class is comparable in response pattern to the fifth latent class in the student sample.

In the third latent class, none of the three interaction terms is significant. Additionally, raters in this class do not seem to substantively employ the scale markers differentially for white versus black candidates. As such, this class of individuals appears to exhibit no DTF. Indeed, the results of the analyses on the adult sample are entirely consistent with those of the conducted on the student sample. First, in each case, the class with no DTF constituted roughly 65% of the sample. Second, as in the student sample, this is the only class wherein which the effects-coded variable for the first step is substantively below zero. That is, this is the class wherein which raters seemed to avoid giving applicants a rating of “1”.

In the fourth latent class, raters are much stricter in their ratings of black applicants when moving from a marker of “1” to a marker of “2”. Conversely, these raters seem to employ the scale very consistently when rating white candidates. Consequently, this latent class seems to be comprised of individuals who exhibit DTF against black applicants, but not in favor of white applicants. Thus, this class is also comparable to one of the classes found in the student data.

In the fifth latent class, raters are much stricter in their ratings of black applicants when moving from a marker of “2” to a marker of “3”. Conversely, raters in this class appear to employ the markers of the scale consistently for white applicants. It appears that individuals in this latent class exhibit DTF against black candidates, but not toward

white candidates. This class directly corresponds to the fourth latent class in the student sample in terms of the patterns of responding.

In sum, five latent classes were uncovered. These classes were comprised of individuals who exhibited varied patterns of DTF: a) DTF against whites and DTF against blacks (class 1), b) DTF against whites but not in favor of blacks (class 2), c) DTF against blacks but not in favor of whites (classes 4 and 5), and d) no DTF (class 3). Of these five classes, four perfectly replicated the latent classes found in the student sample. The remaining class was similar to one found in the student sample, but exhibited marked differences in responses to white candidates when moving from a marker of “1” to a marker of “2”. Thus, at least four of these classes seem to represent stable sub-populations of individuals, at least in the context of the current experimental stimuli.

Hypothesis 4 predicted that rater race, rater prejudice, and ethnic identity would predict latent class membership. I examined the statistical significance of pair-wise comparisons in prediction by these scales. Additionally, I averaged deviations from each group on each measure against all other groups. To this average, I added the mean between-participant score on each scale to create estimated marginal means. These means, and a comparison between the student and adult analyses, are provided in Tables 37 and 38.

When using the scales initially developed in Study 2, there are no significant differences between classes in terms of race. However, there are two trends, such that the second (DTF against whites, not in favor of blacks), and third (no DTF) classes contain more black individuals than the first class (DTF against whites and DTF against blacks). Similarly, when using the new scales to predict class membership, there is a trend such

that the second (DTF against whites, not in favor of blacks) class contains more black individuals than the first class (DTF against whites and DTF against blacks). Overall, when looking at this analysis, classes 1 (DTF against whites and DTF against blacks) and 5 (DTF against blacks but not in favor of whites) have the largest proportion of white individuals, whereas class 2 (DTF against whites, not in favor of blacks) has the largest proportion of black individuals.

Additionally, results of the analyses using the replicated scales reveal that individuals in class 3 (the no DTF class) have a significantly lower ethnic identity than individuals in classes 1 (DTF against whites and DTF against blacks), 2 (DTF against whites, not in favor of blacks), and 5 (DTF against blacks but not in favor of whites). These results are largely replicated with the new scales employed in Study 3, with the exception of the apparent significant difference between class 3 and class 1. Moreover, individuals in class 3 (no DTF) have the lowest ethnic identity, and individuals in classes 1 (DTF against whites and DTF against blacks), 2 (DTF against whites, not in favor of blacks), and 5 (DTF against blacks but not in favor of whites), all exhibit high ethnic identity. Thus, although ethnic identity does appear to be related to DTF, high ethnic identity appears to promote DTF-against the out-group, or applicants of the other race. While this supports the hypothesis for black participants, it does not support the hypothesis that in-group belonging for white participants would result in DTF-for white candidates.

Examination of differences in prejudice between classes yields some interesting results. First, class 3 (class with no DTF) is higher in prejudice against whites than class 5 (DTF against blacks but not in favor of whites), in the replication of Study 2, or class 1

(DTF against whites and DTF against blacks), 2 (DTF against whites, not in favor of blacks), employing both forms of prejudice scales. Additionally, individuals in class 3 (class with no DTF) have higher average prejudice-against-blacks scores than class 2 (DTF against whites, not in favor of blacks), in both analyses. Finally, individuals in classes 3 (class with no DTF) trend toward exhibiting higher levels of prejudice against blacks than individuals in class 4 (DTF against blacks but not in favor of whites). Thus, unexpectedly, a class with the one of the highest levels of prejudice against whites (class 3) is also the class of individuals who exhibits no DTF in their responding patterns. Consequently, there is no support for the hypothesis that prejudice predicts DTF-against.

Finally, I examined patterns of motivation to hide prejudice across classes. First, class 4 (DTF against blacks but not in favor of whites) tends to be more motivated to hide prejudice against blacks than class 3 (no DTF) across both analyses. Additionally, in the analysis using the new prejudice scales, classes 2 (DTF against whites, not in favor of blacks) and 3 (no DTF) trend towards having a higher motivation to hide prejudice against whites than class 1 (DTF against whites and against blacks).

Looking across the student sample and the two analyses of the adults' data, some trends emerge for each class. Specifically, one class is considered to exhibit no DTF. Individuals in this class tend to be high on prejudice against whites and low on ethnic identity. Additionally, both black and white individuals make up this class. Thus, it may be that the secret to exhibiting no DTF is not related to prejudice, but rather, to have a low identification with one's ethnic group. That being said, the form of non-DTF exhibited is not, arguably, the most desirable form. Specifically, while these individuals certainly did not display DTF toward or against either ethnic group, they also employed

the scale in a globally lenient fashion, such that even poor candidates were given fairly high scores. Perhaps this score compression onto the upper end of the scale—rather than individual difference traits—accounts for the apparent no-DTF in these individuals' responses.

Second, one group was found in both student and adult analyses where raters exhibited DTF against whites but not in favor of blacks. Across analyses, this group is primarily comprised on black individuals with a high ethnic identity. These individuals exhibit low prejudice against blacks and low motivation to hide prejudice against either blacks or whites. A similar—but somewhat different—profile was found where individuals exhibited DTF against blacks, but not in favor of whites, at the first step in the Likert scale. These individuals were mostly white, had a low ethnic identity, low levels of self-reported prejudice against either blacks or whites, and high levels of external motivation to hide prejudice against whites and blacks.

Another class also exhibited DTF against blacks, but not in favor of whites. While this group of individuals appeared to exert DTF at a different point in the Likert scale relative to the prior two groups, the profile of this group parallels the group profile for raters who exhibited DTF against white candidates. The individuals in this class were primarily white and had a high ethnic identity. Additionally, they exhibited medium to high levels of prejudice against blacks. Like the prior class of individuals, these participants exhibit a low motivation to hide prejudice against either black or whites. Thus, it may be that prejudice—in concert with ethnic identity and a low motivation to hide prejudice—may foster conditions of DTF-against the out-group. However, it is interesting to note that DTF against blacks does not stem only from prejudice or ethnic

identity, as two entirely different groups of individuals exhibit such DTF in evaluating interview candidates.

Finally, one group did not perfectly replicate across the student and adult samples. This group of individuals displayed a very rich pattern of responding, where they appeared to exhibit DTF against blacks at one point in the Likert scale, and DTF against whites at another point in the scale. This combination was not predicted in the current study's hypotheses. These individuals are high in motivation to hide prejudice against blacks, high in ethnic identity, and overwhelmingly white. Interestingly, their apparent DTF against whites may be accounted for by their desire to mask their prejudice against blacks. However; these white participants' high levels of ethnic identity may still be driving their demonstrated DTF against blacks. Consequently, their final response pattern is as ambivalent and inconsistent as their defining characteristics.

### **Supplementary Analyses**

While the prior analyses illuminate the roles of race, prejudice, and motivation to hide prejudice on rating patterns, it is possible that different demographic characteristics influence patterns of responding. Consequently, I performed a series of chi-square tests and multinomial logistic regressions on both the student and adult samples to assess the extent to which demographic variables impacted latent class membership.

In the student sample, I conducted chi-square tests assessing the relationship between gender, ethnicity, religion, socio-economic status, political orientation, major (psychology or not), guessing the purpose of the study, and latent class membership. The majority of these tests yielded non-significant results. However, the extent to which participants guessed that race was involved significantly related to class membership

( $\chi^2(8) = 23.91, p < 0.05$ ). Specifically, all of the individuals in two classes—those who exhibited DTF against whites, and those who exhibited an inconsistent DTF pattern—guessed that the study was testing race. Additionally, I conducted multinomial regression analyses to assess the extent to which religiosity, age, and GPA impacted latent class membership. None of the analyses are significant. Thus, it appears that demographic variables do not predict patterns of responding in the student sample, whereas awareness of race might.

Next, I conducted a series of chi-square tests on the adult data to assess the extent to which gender, ethnicity, religion, socio-economic status, political orientation, highest level of education, prior experience as a firefighter/EMT, prior experience interviewing, major (business or not), or guessing the purpose of the study predicted latent class membership. The majority of these tests yielded non-significant results. However, highest level of education is associated with class membership ( $\chi^2(16) = 29.44, p < 0.05$ ). Specifically, the latent profile with inconsistent patterns of DTF is associated with lower education, whereas the latent classes depicting DTF against whites and blacks at the first steps are associated with the largest proportion of individuals with graduate-level educations. Additionally, I conducted multinomial regression analyses to assess the extent to which religiosity, age, and college GPA impacted latent class membership. All of these yielded non-significant results. Thus, for adults, education may predict response patterns. However, a key issue to notice is that higher education does not result in less DTF. For a summary of the results of the analyses on the student and adult data, refer to Table 39 and Table 40.

## Discussion

While discrimination in employment scenarios remains a salient concern, it is not always clear what drives it. Indeed, a score discrepancy indicative of discrimination may stem from DTF-for a relevant in-group, from DTF-against some disliked group, or from both. Moreover, different forms of discrimination may be driven by different personal characteristics, such as in-group belonging and prejudice.

The current studies sought to investigate the usefulness of IRT to examine DTF, and LCMM to assess the extent to which different “types” of DTF could be predicted by individual differences. Specifically, I assessed DTF-for and DTF-against across three studies. First, I assessed the extent to which raters were more lenient toward same-race applicants in a field study, in which trained raters evaluated hundreds of candidates for an entry-level firefighting position. Second, I conducted both IRT and LCMM analyses in two experiments to examine the extent to which raters could be classified into five latent classes with different combinations of DTF-for, DTF-against, and no DTF. In particular, I expected that five classes would emerge, such that the following labels could be employed to describe the response patterns within these classes: a) no DTF, b) DTF-against blacks and not for whites, c) DTF-against blacks and DTF-for whites, d) DTF-for whites and not against blacks, e) DTF-against whites and not toward blacks. Finally, I predicted these classes using ethnic identity (Study 3), rater race, prejudice, and motivation to hide prejudice (Studies 2 and 3). Specifically, I expected that in-group belonging would predict DTF-for a favored in-group (for whites), or against an out-group (for blacks), whereas prejudice would exclusively predict DTF-against individuals from a disliked group.



First, the field study provided initial support, via the significant rater race and applicant race interaction, for the hypothesis that individuals are more lenient toward applicants of their own race. More detailed support for this hypothesis was found in the experimental studies, however. In both experimental studies, raters had an easier time moving from a “1” to a “2” when assessing candidates of their own race. In other words, black raters are more lenient toward black applicants, relative to white applicants, and white applicants are rated more leniently by white raters relative to black raters.

Another interesting finding emerged in addition to the differential support for Hypothesis 1 derived from the field and experimental studies. Specifically, raters in the field sample displayed much more consistent usage of the Likert scale relative to both experimental samples. That is, field raters had lower average difficulty parameters than either experimental sample. Given the vast differences in training between the field raters and the experimental raters, it is possible that both the apparent same-race preference and the difference in usage of the scale is a function of differential amounts of rater training. Additionally, these differences might have emerged as a function of applicant characteristics. Specifically, field raters evaluated a large variety of applicants who exhibited natural variation on competence. In contrast, experimental raters evaluated six applicants who were specifically picked to maximize competence dispersion. Consequently, these portrayed applicants did not represent all possible competency levels—instead, they represented extremely low, extremely high, and perfectly average. It is possible that scale usage in the experimental sample was less consistent because no simply below average or simply above average candidates were portrayed.

In each of the experimental samples, I found that five latent classes fit the data best. Four of these classes were replicated across the student and adult samples. Specifically, these four classes described individuals who exhibited: a) DTF against whites but not in favor of blacks, b) DTF against blacks but not in favor of whites (two classes), and d) no DTF. The remaining class of respondents varied somewhat between student and adult samples. In both samples, participants were strict toward black candidates when moving from a “1” to a “2”, and toward white candidates when moving from a “2” to a “3”. However, in the student sample, participants were also strict toward white candidates when moving from a “1” to a “2”. This difference led to the final class receiving different labels for the student (DTF against whites and somewhat in favor of blacks) and adult (DTF against whites and against blacks) samples. Despite the inconsistencies between student and adult results in this final class, three of the five classes were predicted in Hypothesis 2 (no DTF, DTF against whites but not in favor of blacks, DTF against blacks but not in favor of whites). Notably, the two hypothesized classes that did not receive empirical support were those which expected DTF-for (DTF for whites and against blacks, DTF for whites but not towards blacks).

Additionally, the individual differences which predicted latent class membership were not those forwarded in Hypotheses 3 and 4. Specifically, the group with no DTF was composed of black and white individuals with high prejudice against whites and low ethnic identity. Moreover, the group that exhibited DTF against whites but not in favor of blacks was comprised primarily of black participants who report low prejudice against blacks, a low motivation to hide prejudice, and a high ethnic identity. Similarly, one group that exhibited DTF against blacks but not in favor of whites was composed of

white individuals who reported medium to high prejudice against blacks, a low motivation to hide prejudice, and a high ethnic identity.

The DTF pattern in this latter group was unique, in that it was very “hard” for black applicants to be rated a “3”, but when applicants did get to that marker, it was very “easy” for black applicants to receive a “4”. In other words, these raters polarized the scale, such that black applicants were rarely given a “3”, and were rather classified as either “good” or “bad”. Indeed, the prior literature emphasizes how discrimination may emerge through polarized and homogenous responses for individuals belonging to a given sub-group (e.g. Alvesson and Billing, 1992). That is, individuals confronted with information that contrasts with their prior expectations tend to ignore this information until it becomes overwhelming (e.g. Foti, Knee, & Brackert, 2008; Hanges, Braverman, & Rentsch, 1991; Nowak et al., 2000). As such, prior to the point where the information cannot be ignored, disconfirming information remains un-integrated in preference to the prior expectation. After this point, individuals suddenly and dramatically change their opinion (Foti et al., 2008; Hanges et al, 1991), instead of gradually adapting it, thus leading to polarized responses. Consequently, raters in this latter group display a very “classic” pattern in intergroup bias.

Additionally, across the three groups previously discussed, ethnic identity (and hence, in-group belonging) played a large role in predicting DTF-against versus no DTF. As previously discussed, the intergroup bias literature implies that discrimination stems primarily from same-race favoritism rather than different-race derogation (Brewer, 1979; 1999; Brown, 2000). However, it is apparent that the discrimination uncovered in the current studies stems from derogation, rather than favoritism. Moreover, race and ethnic

identity appear to most strongly predict DTF-against, rather than DTF-for. Indeed, the literature on intergroup bias also suggests that conditions of threat—such as those associated with competition—inspire out-group derogation (Hewstone et al., 2002). It may be the case that the current context, where presumably few applicants would be hired, provided the appearance of competition between black and white candidates.

The second group that displayed DTF against blacks but not in favor of whites was composed of both black and white individuals with a low ethnic identity, low levels of reported prejudice, and high levels of motivation to hide prejudice. According to the theory of flexible corrections (Petty & Wegener, 1993; Wegener & Petty, 1995), individuals who are aware of their prejudices may seek to compensate for them by changing their responses. These individuals, despite their motivation to mask prejudice, do not admit to actually having high levels of prejudice. Consequently, it may be that these individuals hold implicit prejudices against blacks of which they are not aware.

Last, the remaining group—which did not perfectly replicate across student and adult samples—displayed an unusual pattern of responding. Specifically, these individuals may appear to exhibit DTF against both blacks and whites. Indeed, this group is comprised of white individuals who are high in a motivation to hide prejudice against blacks and high in ethnic identity. Their in-group belonging may be driving their apparent discrimination against black applicants, whereas their motivation to appear non-prejudiced toward blacks may be driving their reactions to white candidates.

### **Limitations and Future Directions**

While the results of the current research are promising, there were a number of limitations in the design of the experimental studies that impact the generalizability of

their findings. Several limitations directly address the content of the interviewee scripts. First, it is important to note that the script from the medium-scoring black candidate was rated consistently higher than the script from the medium-scoring white candidate by participants in both experiments, regardless of the race of the candidate presented to participants. However, in the field, the black candidate from whom this script was derived received a rating of “3” across these situations, and the white candidate from whom the other script was derived received an average rating of “3.33” on these situations.

This apparent reversal may indicate one of two things. First, race may matter in terms of content of the interview. That is, due to differences in background and experiences, black and white candidates may discuss different things, approach problems from different perspectives, or communicate their intentions differently. Secondly, this apparent reversal may be indicative of discrimination in interview ratings in the field. In other words, it is possible that the black applicant truly gave a better response than the white applicant. However, DTF in the rating process in the field may have resulted in seemingly depressed scores for this applicant. If this is the case, the conclusions of the current study may be somewhat suspect. In other words, the scripts were deliberately selected to provide “objective” poor, medium, and high-quality candidates. To the extent that the field ratings were truly not objective, conclusions regarding the DTF of participants derived from the differential usage of scale points may be less accurate.

Additionally, beyond the issue of the race of the candidate providing the script, the current study only utilized six scripts. In general, future research should explore a broad array of potential applicants in order to tease apart issues of how script content,

candidate language, and other distinguishing characteristics impact ratings. Moreover, researchers may also wish to develop their own scripts from the rating criteria rather than relying solely on actual applicant responses.

Another limitation of the experiments relates to the audio recordings of the scripts. As previously discussed, six white male actors provided the audio for the current study. Despite attempts to make their voices more racially ambiguous, participants still responded differently to the vocal profiles. Specifically, candidates that sounded black were rated more poorly than candidates who sounded white. Moreover, this relationship varied as a function of the apparent race of the applicant in the student sample—white-sounding black applicants were rated favorably, whereas black-sounding white applicants were not. Consequently, the vocal profile of the candidates may have influenced ratings as much, or in conjunction with, discrimination triggered by photographs of candidates. Certainly, this is an issue that should be explored in greater depth in future research. Specifically, researchers may deliberately manipulate and explore the effects of vocal profiles on ratings.

Similarly, researchers may wish to systematically evaluate how different methods of presentation (e.g. video, audio, text) impact ratings. Indeed, previous research shows that communicator characteristics—such as those which might prompt discrimination—impact evaluations more strongly when presented via audio and videotape rather than in text (Chaiken & Eagly, 1983; Ziegler, Arnold, & Diehl, 2007). Possibly, providing text allows raters to centrally process the information and actively seek out counter-stereotypic statements. In audio and video, however, information must be processed more quickly, which might prevent active searching for counter-stereotypic associations. A

better understanding of how the presentation impacts ratings will allow for practitioners to construct interviews and evaluate interview responses in the most accurate form.

Other limitations of the current research concern the directions provided in the experiment. First, participants were told that they were welcome to take notes at their own discretion. As discussed, the types of notes different participants took varied widely. That is, some participants took no notes at all, others took notes about competency judgments, and still others took extensive behavioral notes. Likely, the type of notes taken by participants would affect DTF. Indeed, some recent research indicates that structured recall of behavior reduces discrimination against women and minorities (Baltes, Bauer, & Frensch, 2007; Bauer & Baltes, 2002). Consequently, it seems reasonable that behavioral notes might have a similar dampening effect. Future research should thus consider manipulating instructions such that participants are allowed to take specific types of notes in order to understand how note-taking impacts DTF. Indeed, a difference in training may explain why the average difficulty parameters in the field sample were so much lower than those in either experimental sample.

Further, while the findings of the latent class mixture model analysis replicated across two samples, it is still very likely that there is some degree of sample bias across the two experiments. First, while there is some age diversity and socio-economic status diversity in the adult sample, participants in both experimental samples had limited experience in firefighting, EMT, or emergency services positions, and limited experience interviewing. Indeed, as discussed previously, there were large differences in the absolute difficulty parameters between experimental and field raters. These differences might be attributed to differences in the amount of experience between these sets of raters.

Additionally, all experimental participants were receiving some form of compensation for completing the study. Specifically, student participants received extra credit for participating, and adult participants received one dollar. Certainly, the motivations of individuals who complete studies for extra credit, or for a small amount of compensation, would vary from the motivations of those whom would only complete such studies for higher payment, or for free. Moreover, the study was run entirely over the internet, essentially ruling out individuals without easy access to internet (lower socio-economic status, in particular) as participants. Consequently, the findings of the experimental studies, while promising, may have limited generalizability, as they were drawn from a limited sample of the overall population.

Both experimental classes also had relatively conservative sample sizes for five classes, given that one class constituted roughly 65% of each sample. That is, only 10 or so participants would be classified as part of classes constituting 5% of the overall sample. With larger samples, it is possible that some of the classes would “split”, resulting in more than five latent classes, or that class profiles would differ. Indeed, limitations in scripts, instructions, and samples might have affected both the latent class profiles and what predicted membership in latent classes. Thus, researchers should systematically assess these potential confounds in order to better understand how DTF manifests in interview contexts.

Finally, the field sample (18 raters) was too small for accurate comparisons against the experimental sample. Consequently, it is difficult to compare and assess the differences between the experimental and field samples. Future research should strive to



find larger samples of more experienced and highly trained raters. Some samples would provide a more generalizable test of DTF in actual interview evaluations.

### **Theoretical Implications**

The current study offers profound implications for both the study of discrimination, and for understanding, and hopefully preventing it, in the workplace. First, discrimination is not necessarily a linear phenomenon. That is, most prior studies have concerned themselves primarily with means and variances. Consequently, many studies have been relatively one-dimensional, and have not been able to capture the complex nature of DTF or its drivers. That is, a linear regression analyses would have identified general trends in responding across the entire sample. In contrast, combining IRT and latent class modeling provides a nuanced view of DTF and its drivers. Specifically, I was able to identify potential sub-populations from which the samples were drawn, as well as the rich patterns of individual differences that predict different responding styles. Indeed, many of the individual difference combinations I was able to assess in the current study would not be detectable through linear analyses.

The current studies sought to identify when DTF-for and DTF-against would manifest in interview evaluations. However, only DTF-against was found, not DTF-for. Perhaps the lack of support for DTF-for is, as previously discussed, due to intergroup threat being introduced (Hewstone et al., 2002). Indeed, if this is the case, future research may wish to investigate DTF in less competitive workplace scenarios. Perhaps, for example, evaluations and assessment are somewhat less competitive relative to selection. Further research in such areas may reveal that DTF-for a given group does indeed operate in the workplace—just, perhaps, not during selection.

One particularly interesting finding in the current study is that two groups of individuals appeared to exhibit DTF against blacks but not in favor of whites. Importantly, DTF manifests at different points in the Likert scale, and different “types” of individuals exhibit different types of DTF. It appears that direct discrimination toward individuals from an out-group is exhibited between Likert markers 2 and 3, and is a product of a high ethnic identity and relevant prejudicial beliefs. Conversely, a more ambivalent DTF against blacks appears to manifest between Likert markers 1 and 2, and is driven by a motivation to hide prejudice against blacks. To date, discrimination driven by derogation has been treated as a single phenomenon. The results of the current study suggest that derogation may be multifaceted and complex. Consequently, it may be important to further explore different forms of DTF-against. Perhaps different motivations relate to different forms of DTF, as well as different correlates and drivers.

### **Practical Implications**

In addition to the implications for research provided by the current findings, the work discussed here also suggests a number of implications in terms of interventions to reduce discrimination in interview evaluations as well as training and selection of raters. First, desired interventions may vary greatly depending on the nature of the DTF exhibited. For example, given that DTF-against was most apparent in the current study, it may be important to focus intervention efforts on lessening the salience of ethnic identity. Specifically, a low ethnic identity contributed to no DTF, even in the presence of high levels of prejudice. Thus, it may simply be important to reduce the salience of race in hiring, so as to reduce the likelihood of ethnic identity impacting evaluations. Further, as discussed, DTF appears to largely be a function of race and ethnic identity. Consequently,

ensuring that individuals of different races rate each candidate (e.g. one black rater, one white rater) may help balance potential intergroup bias, and hence, may combat overall discrimination in a given candidate's evaluation.

Additionally, as discussed, two different forms of DTF against blacks were manifest in the current research. Consequently, different forms of interventions might be necessary to combat direct versus ambivalent drivers of discrimination. Indeed, companies may wish to employ multiple interventions when training raters in order to better address all relevant drivers of differential responding. In addition to training interventions, it may be prudent for organizations to select individuals who have low ethnic identity and a low motivation to hide prejudice, given that both seem to drive DTF.

Indeed, organizations can also assess DTF in evaluations through a closer examination of candidate ratings. Specifically, usage of the Likert scale for black and white applicants, and by rater, can be assessed simply by mapping out the frequencies of usage of each scale marker (e.g. "1", "2", "3", etc.). The extent that the scale markers are used with a different frequency for white and black applicants can provide organizations a visualization of potential DTF, and thus, discrimination, in evaluations. Here, however, it is imperative to consider the pool from which applicants are being drawn to ensure that differences in usage of the Likert scale are indicative of DTF rather than reflections of true differences in ability.

Finally, one interesting finding in the current research is that DTF appears to be a function of the lowest point on the Likert scale. Prior research, which has focused on means rather than usage of particular scale points, has not been able to assess at what point in a Likert scale differential ratings of applicants occurs. The current research

suggests that DTF appears to manifest at the lower end of the scale, when raters are decided if a candidate is qualified enough to be rated as a “2” or a “3”. Thus, such raters may see candidates belonging to a particular group as either “poor” or “average”, rather than being able to distinguish between different levels of ability at lower levels of competency. If these findings generalize across samples and contexts, discrimination may not be as serious a concern in the field as previously expected. Specifically, interview candidates will not move forward in the selection process unless they are rated highly. Thus, as long as raters employ the upper portions of the scale equivalently (e.g. no DIF on steps 3 or 4), any DTF in the lower portion of the scale may not translate into different hiring decisions. Consequently, more work should be done in this area to determine where on the scale DTF, and thus, discrimination, may manifest across contexts.

### **Conclusion**

In sum, the current research sought to investigate racially motivated DTF in a hiring context. Specifically, I investigated DTF-for, DTF-against, and no DTF across three studies (one in the field, two experimental). Moreover, I assessed the latent class membership of raters as well as individual difference factors that predict latent class membership. Some support for hypotheses—namely, those investigating same-race favoritism (IRT) and latent class response profiles (LCMM)—was found. However, DTF-for was not apparent in the current data, in –group belonging predicted DTF-against, and motivation to hide prejudice drove discriminatory responding at least as much as prejudice itself did. Overall, I obtained strong support for the application of IRT to assessing DTF. Future research should continue to explore the application of this analytical technique within the realm of research on discrimination in evaluations.

## Appendix A

### Situation #1

Assume that you are an entry-level firefighter. You work on a 24 hour shift. During the shift, you and your co-workers are required to work and live closely together. Assume that one of your co-workers displays behavior that you find irritating. For example, he makes slurping noise when drinking and changes the TV channel without asking others.

#### Follow-Up Questions

What actions, if any, would you take in this situation and why? Would you say anything to this co-worker? Why or why not? What would you say? Would you involve anyone else? If so, who and why? If not, why not?

	<b>Benchmarks</b>
	<b>Outstanding = 5</b>
	Candidate identifies several alternative methods for effectively dealing with situation. For example, s/he would discuss issue directly with person (if behavior is under control of person) <b>OR</b> would tolerate behavior if not under person's control
	<b>More Than Acceptable = 4</b>
	Discusses the issue directly with co-worker (candidate does not shy away from addressing the issue).
	Let co-worker know in a tactful way that his/her behavior is annoying
	<b>Acceptable = 3</b>
	Candidate understands that s/he might not be able to change the co-worker
	Candidate tolerates behavior if behavior is seen as not being under the person's control
	Requests advice from a peer on how to handle co-worker.
	<b>Less Than Acceptable = 2</b>
	Tells the co-worker to stop doing it
	<b>Unacceptable = 1</b>
	Complains to supervisor before talking to co-worker
	Complains to other co-workers

Candidate ID		Preliminary Rating	1   2   3   4   5 (Please circle your rating)	Final Rating	1   2   3   4   5 (Please circle your rating)
Assessor Number					

## Situation #2

Imagine that you and another person are newly hired firefighters. You both work at the same station on the same shift. One day the Lieutenant assigns both of you the task of waxing the fire truck. The Lieutenant tells you that the job has to be completed in one hour. The Lieutenant will be back to inspect the job. You start waxing one side of the truck and the other person starts waxing the other side. After about 40 minutes you are finished with your side and you see that the other person is not finished. In fact, he is far from finished.

Follow-Up Questions

What actions, if any, would you take and why? Would you say anything to the other person waxing the truck? If so, what? If not, why not?

Would you say anything to anyone else? If so, what? If not, why not?

	<b>Benchmarks</b>
	<b>Outstanding = 5</b>
	Candidate develops plan about how to break the task into different parts so they can work together to complete task by deadline <b>AND</b> Suggests to person that they can finish the work on time if they work together
	<b>More Than Acceptable = 4</b>
	Offers to help the person
	Candidate says "let me help you finish because it is close to time" to the person
	<b>Acceptable = 3</b>
	Would not say anything to anyone else, because candidate believes that situation should be resolved between candidate and person
	Candidate offers suggestions about how the person can speed up his/her work
	<b>Less Than Acceptable = 2</b>
	Reminds person of deadline but does nothing else
	Candidate indicates that s/he finished her/his side and so s/he is done
	<b>Unacceptable = 1</b>
	Assigns blame to the person for the task not being completed
	Would ridicule the person

Candidate ID		Preliminary Rating	1   2   3   4   5 (Please circle your rating)	Final Rating	1   2   3   4   5 (Please circle your rating)
Assessor Number					

## Situation #3

Imagine that you are a new firefighter. You have just graduated from the firefighter academy. You know that even though you have graduated from the academy, you have only learned the basics of being a firefighter. There is still much to learn. You know how to perform certain tasks, but you still hesitate while performing them. You are not as proficient as you need to be. In addition to improving your current skills, you also realize that you have to learn new skills and information. For example, you have to learn the fire station's territory as well as continually updating your knowledge of procedures.

Follow-Up Questions

What would you do to learn the firefighter job once you are at the station?

<b>Benchmarks</b>
<b>Outstanding = 5</b>
Would talk to more experienced firefighters/officers to clarify what it takes to be successful
Identifies multiple resources that s/he could use to learn the job. For example, uses manual as a learning tool; asks supervisor for specific feedback on her/his performance.
<b>More Than Acceptable = 4</b>
Candidate describes a detailed set of systematic steps to learn the job
Candidate indicates that s/he would actively seek out opportunities to broaden/practice skills
<b>Acceptable = 3</b>
Candidate indicates that s/he would observe other firefighters at station and follow their example
Would try to learn as things occur
<b>Less Than Acceptable = 2</b>
Candidate indicates that s/he would depend on memory as main method of learning job
<b>Unacceptable = 1</b>
Candidate cannot identify any resources to use to learn the job
Depends on others to make sure that s/he learned job. Takes no responsibility for own learning

Candidate ID		Preliminary Rating	1   2   3   4   5 (Please circle your rating)	Final Rating	1   2   3   4   5 (Please circle your rating)
Assessor Number					

#### Situation #4

Assume that you get the firefighter job and you are now working at a station somewhere in Jefferson County. It is late at night and you are about to go to bed. A civilian knocks on the station's door. You answer it. The civilian at the door has come to the station before- during lunch and supper times, asking to have his blood pressure checked. Many firefighters, including you, have checked his blood pressure for him. He is currently argumentative and appears to be anxious. Once again, he wants his blood pressure checked

#### Follow-Up Questions

What actions, if any, would you take and why? Would you say anything to the civilian? If so, what? If not, why not? Would you say anything to anyone else? If so, who and why? If not, why not?

	<b>Benchmarks</b>
	<b>Outstanding = 5</b>
	Candidate's response indicates that s/he recognizes that firefighters are on duty 24 hours a day
	<b>More Than Acceptable = 4</b>
	Candidate recognizes the need for candidate to maintain composure – remains calm
	Talks calmly to person while checking blood pressure
	Indicates that s/he would talk to the candidate courteously
	<b>Acceptable = 3</b>
	Would let the civilian vent until candidate can begin to control the conversation
	Says something to civilian to calm civilian. For example, "please clam down" or "I'm trying to help"
	Would provide general information to civilian about target blood pressure reading
	<b>Less Than Acceptable = 2</b>
	Candidate fails to see any real concern with situation (blood pressure problems), beyond the behavior of the civilian
	Candidate indicates that s/he does not have to take that kind of behavior from civilian
	<b>Unacceptable = 1</b>
	Takes minor retaliation against civilian. For example, pumps blood pressure cuff tighter than it needs to be
	Refuses to take blood pressure
	Responds back at the same emotional level as civilian
	Would ask someone else to take civilian's blood pressure

Candidate ID		Preliminary Rating	1   2   3   4   5 (Please circle your rating)	Final Rating	1   2   3   4   5 (Please circle your rating)
Assessor Number					



### Situation #5

The job of being a firefighter can be very stressful. The work can be dangerous and demanding. Interacting with victims can be emotionally draining. Describe a very stressful situation that you have encountered in the past. The stressful situation could be work related but it doesn't have to be. You don't have to describe the situation in detail but provide general information so that it is clear what made the situation stressful to you.

### Follow-Up Questions

What did you do to deal with the stress? Did you talk to anyone about this stressful situation? If so, who and why? If not, why not?

	<b>Benchmarks</b>
	<b>Outstanding = 5</b>
	Describes multiple methods of relieving stress. Methods include some actions taken while on the job and others while off the job. For example, taking breaks during work, exercising away from work
	<b>More Than Acceptable = 4</b>
	Analyzed situation to figure out solution to stressful situation
	Has experience relieving stress using healthful methods. For example, exercise, meditation, etc.
	<b>Acceptable = 3</b>
	Candidate can identify at least one concrete example of a method of reducing stress
	Provides an example with a moderate level of stress. Identifies a reasonable approach for handling the stress.
	<b>Less Than Acceptable = 2</b>
	Candidate cannot provide any <i>specific</i> examples of how s/he has handled stress
	Says "Stress doesn't affect me" or similar denial
	Candidate didn't talk to anyone about stress
	<b>Unacceptable = 1</b>
	Lost self-control in critical situations
	Became argumentative with co-workers
	Assigned blame to others for stress
	Sought relief through unhealthy methods (e.g., drinking and/or drugs)

Candidate ID		Preliminary Rating	1   2   3   4   5 (Please circle your rating)	Final Rating	1   2   3   4   5 (Please circle your rating)
Assessor Number					

## Appendix B

Good morning. Are you here for the study, “Interview Assessment?” (Wait for assent). Please sign in on this sheet (usual sign-in sheet). Thank you. I’d like you to sit here, please (Direct to appropriate computer). Please read over this consent form (hand to participant), and let me know if you have any questions. If you feel comfortable with it, and have read it over, please sign the consent form and initial each page on the upper right-hand corner. Let me know when you are ready to begin.

Great, thank you (put consent form in drawer). The study you are participating in today is designed to help us understand how people rate job candidates. Specifically, we are focusing on how people assess interview responses. Thus, you will be taking part in what is called a “work simulation”, which is designed to mimic decisions and actions that are commonly made in organizations. This specific work simulation requires that you act as an interview assessor for six candidates applying for an entry-level firefighting position. Each of these six candidates will respond to the same three questions. One of these questions is about how the candidate would handle a situation where a coworker is not pitching in to do his or her fair share, the second is about how the candidate would handle a situation where group work is required and his or her coworkers are not doing their work as quickly or as thoroughly as the candidate, and the final situation is about how the candidate would respond to a civilian interruption at 2 a.m. For each candidate, you will listen to and rate his or her response to each individual question, and then provide an overall rating of that candidate as a whole. While you are listening to each candidate’s response, you will have a space on the computer to take notes on what he or she is saying. Finally, you will be asked to rank the candidates at the conclusion of rating all six separately. To facilitate the ranking, you are welcome to jot down notes and your overall rating of each presented candidate on the sheet of paper provided.

Your first page provides you with detailed information on each of these situations. Additionally, you are provided with “benchmarks” that should help you rate each candidate. These benchmarks provide information on responses that might be characteristic of an “outstanding”, “acceptable”, or “unacceptable” response. It is important to note that candidates do not have to do all of the actions listed under any given anchor to get that rating. Indeed, a candidate might do none of the actions listed. Instead, these benchmarks are provided to give you a feel for the kinds of things that the fire house is looking for in candidates for this firefighting position. They are not set in stone; use your best judgment to compare the candidates’ behaviors against these benchmarks in order to determine a final rating. Remember to try to be objective as possible when rating each candidate.

Please read over the situations and benchmarks in detail and let me know if you have any questions. Otherwise, you may proceed to the next page by clicking “next” at the bottom of the screen to begin rating candidates.

(If they have no questions, or once they have been answered). Great. Let me know when you are finished with this portion of the study.

## Appendix C

In the following exercise, you will be assessing interviews from six candidates. These candidates are applying for an entry-level firefighting job. In their interviews, each candidate responded to the same three questions. As an assessor, you will be asked to rate each candidate on his or her answers to each question, and then provide an overall rating for each candidate. Please try to be objective as possible when rating the candidates.

During this exercise, you will view the question text and some information about each candidate. You will then listen to his or her response to each question. Space will be provided for you to take notes while you assess these candidates. At the conclusion of each candidate's response to each question, you will be asked to provide a rating of that candidate's response, based on the benchmarks provided. **Please note** that the benchmarks provided for each scale anchor are examples of what might constitute an outstanding, mediocre, or poor response. Candidates are not required or expected to do all of the behaviors listed under any given anchor (e.g. 4 or 5) to get that rating. Indeed, they don't have to do any of the listed behaviors. Instead, these anchors are designed to give you a feel for what the fire department wants a 4 or a 3 to be. As such, your job as a rater is to consider what each candidate said and to compare where each candidate's answer falls compared to the examples given.

At the conclusion of all three responses, you will be asked to rate the candidate's overall performance in the interview. This procedure will be repeated for each of the six candidates.

Please familiarize yourself with the situations, questions, and scoring guidelines before proceeding. The three situations, resulting questions, and their scoring guidelines, are as follows:

### **Situation #1**

Assume that you are an entry-level firefighter. You work on a 24 hour shift. During the shift, you and your co-workers are required to work and live closely together. After you eat dinner with your coworkers each night, everyone who did not help cook the meal is supposed to help clean up the dirty dishes and the kitchen. One of the other entry-level firefighters who works on your shift seems to always avoid cleaning anything by staying seated at the kitchen table until all of the cleaning has been done. You like this co-worker. You realize, however, that he is not doing his fair share of cleaning and he is beginning to irritate you and others.

#### Follow-Up Questions

What actions, if any, would you take in this situation and why? Would you say anything to this co-worker? Why or why not? What would you say? Would you involve anyone else? If so, who and why? If not, why not?

**Benchmarks****Outstanding = 5**

Candidate identifies several alternative methods for effectively dealing with the situation. For example, he/she would discuss the issue directly with the co-worker AND would seek advice from peers/supervisor

**More Than Acceptable = 4**

Candidate states he/she would speak with the co-worker in private in a tactful manner about how the behavior may be bothering others who are helping cook and clean in the kitchen

Candidate emphasizes the importance of talking to the co-worker in a tactful manner about the issue

**Acceptable = 3**

Candidate states he/she would ask for a supervisor's assistance with the co-worker before talking directly to the co-worker

Candidate states he/she would privately explain to the co-worker that he is not doing his fair share of cleaning

Candidate states he/she would ask a more senior firefighter to handle the situation

**Less Than Acceptable = 2**

Candidate states that he/she may not be able to change the co-worker

Candidate states he/she would not talk to anyone else about it because it is a private matter

Candidate states he/she would confront the co-worker about the situation in front of other firefighters

Candidate states he/she would leave a note for the co-worker asking him start doing his part of the cleaning after meals

**Unacceptable = 1**

Candidate states he/she would make fun of the co-worker in front of other firefighters

Candidate states he/she would retaliate in secret against the co-worker

Candidate states he/she would threaten the co-worker with retaliation if he does not start doing his part of the cleaning

**Situation #2**

Assume you are a firefighter. Every six months your station takes a day to paint all of its equipment. All of the firefighters at the station are assigned different pieces of equipment to paint that day. Although you are one of the most experienced painters, you are assigned one of the easiest jobs because of your seniority. After half an hour of work, you are finished painting your equipment. You see that some of the other firefighters are painting their equipment incorrectly and still have a lot left to paint.

**Follow-Up Questions**

What actions, if any, would you take in this situation and why? Would you say anything to anyone? Why or why not? What would you say?

Would you involve anyone else? If so, who and why? If not, why not?

### **Benchmarks**

#### **Outstanding = 5**

Candidate provides several alternative solutions to effectively deal with the situation. For example, the candidate would ask the firefighters in a friendly manner “Can I help you?” AND would show the firefighters how to paint the equipment correctly AND asks a few coworkers to join him/her in helping the slow firefighters

#### **More Than Acceptable = 4**

Candidate states he/she would ask the firefighters in a friendly manner “Can I help you?”

Candidate states he/she would offer help to the firefighters

#### **Acceptable = 3**

Candidate states he/she would ask a few co-workers to join him/her in helping the firefighters

Candidate states he/she would show the firefighters how to do the task more quickly and efficiently

Candidate states he/she would try to motivate the firefighters to work faster

Candidate states he/she would say to the firefighters “let me help you so that we can all finish”

#### **Less Than Acceptable = 2**

Candidate states he/she would say nothing to the firefighters

Candidate states he/she would start to help the firefighters without saying anything to them

Candidate states he/she would wait to see if anyone else helps the firefighters before he/she would help them

Candidate states he/she would tell them how to paint the equipment, but would not show them

#### **Unacceptable = 1**

Candidate states he/she would ridicule the firefighters

### **Situation #3**

Imagine that you are a firefighter at a city within Jefferson County. It is 2:00 in the morning and you have just gotten back from the third call of the night. As you are storing your equipment, a man knocks on the station’s front door. You realize the man is someone you see around the station frequently and you suspect he is living on the streets. When you answer, he says he has an infected finger and is in a lot of pain. You look at his finger and cannot see anything wrong with it. The man demands that you give him some pain killers immediately to help with the pain. When you state that you are not allowed to dispense drugs, he begins to yell insults at you.

#### Follow-Up Questions

What actions, if any, would you take in this situation and why? Would you say anything else to the civilian? If so, what? If not, why not?

Would you say anything to anyone else? If so, who and why? If not, why not?

### **Benchmarks**

#### **Outstanding = 5**

Candidate states he/she would do several things that would effectively handle the situation. For example, he/she would first calm the civilian AND offer to take the civilian to the hospital AND treat the civilian with respect to maintain a good relationship with the public

#### **More Than Acceptable = 4**

Candidate states he/she would tell the civilian she/he will get the supervisor to talk to him

Candidate indicates the need to maintain composure – remain calm

Candidate states he/she would try to be polite to the civilian despite the civilian's behavior

#### **Acceptable = 3**

Candidate states that for safety and witness purposes, he/she would get another firefighter before interacting with the civilian

Candidate states he/she would say things to the civilian to calm him down (e.g., “please calm down”, “let me look at your finger”, etc.)

Candidate indicates that he/she would get help from someone else after trying to calm the civilian him/herself

Candidate states he/she would ask the civilian if he wanted to go to the hospital

#### **Less Than Acceptable = 2**

Candidate indicates that he/she does not have to take that kind of behavior from a civilian

Candidate states he/she would let the civilian vent until the candidate could control the conversation

Candidate states he/she would tell the civilian to leave the fire station

#### **Unacceptable = 1**

Candidate fails to see any real concern with the situation

Candidate states that he/she would ignore the civilian

Candidate states he/she would become argumentative with the civilian

Candidate states he/she would give the civilian the pain reliever drugs

## **Appendix D**

### Prejudice against Blacks

1. Blacks are responsible for creating the racial tension that exists in the United States
2. Discrimination against blacks in the United States today limits their changes to get ahead.
3. Over the past few years, blacks have gotten less than they deserve.
4. Over the past few years, blacks have gotten more economically than they deserve.
5. Over the past few years, the government and news media have shown more respect to blacks than they deserve.
6. Blacks should not push themselves where they're not wanted.
7. It is easy to understand the anger of black people in America.
8. I think that black people look more similar to each other than white people do.
9. If a black were put in charge of me, I would not mind taking advice and direction from him or her.
10. It is likely that blacks will bring violence to neighborhoods when they move in.
11. I get very upset when I hear a white make a prejudicial remark about a black.
12. Some blacks are so touchy about race that it is difficult to get along with them.

**Appendix D, Cont.**Prejudice against Whites

1. Whites are responsible for creating the racial tension that exists in the United States
2. Reverse discrimination against whites in the United States today limits their changes to get ahead.
3. Over the past few years, whites have gotten less than they deserve.
4. Over the past few years, whites have gotten more economically than they deserve.
5. Over the past few years, the government and news media have shown more respect to whites than they deserve.
6. Whites should not push themselves where they're not wanted.
7. It is easy to understand the anger of white people in America.
8. I think that white people look more similar to each other than black people do.
9. If a white were put in charge of me, I would not mind taking advice and direction from him or her.
10. It is likely that whites will bring violence to neighborhoods when they move in.
11. I get very upset when I hear a black make a prejudicial remark about a white.
12. Some whites are so touchy about race that it is difficult to get along with them.



## **Appendix E**

### Motivation to Hide Prejudice – Whites

1. I try to hide any negative thoughts about white people in order to avoid negative reactions from others.
2. If I acted prejudiced toward white people, I would be concerned that others would be angry with me.
3. I attempt to appear non-prejudiced toward white people in order to avoid disapproval from others
4. I try to act non-prejudiced toward white people because of pressure from others.

### Motivation to Hide Prejudice – Blacks

1. I try to hide any negative thoughts about white people in order to avoid negative reactions from others.
2. If I acted prejudiced toward white people, I would be concerned that others would be angry with me.
3. I attempt to appear non-prejudiced toward white people in order to avoid disapproval from others
4. I try to act non-prejudiced toward white people because of pressure from others.

**Appendix F**

What is your gender?

Male      Female

2. Which of the following best describes your race or ethnicity?

Circle whatever identity applies:

African or African American  
Asian or Asian American  
Hispanic

Caucasian / White  
Other: \_\_\_\_\_

How old are you? \_\_\_\_\_

What is your religion?

Christian	Jewish	Buddhist	Islamic
Hindu	Agnostic	Atheist	Spiritual but not Religious
Other: _____			

How religious do you consider yourself to be?

Not very      Slightly      Somewhat      Moderately      Very much

What is your socio-economic status?

Lower                  Middle                  Upper

Is English your native language?

Yes      No

What is your country of origin?

Name of country: \_\_\_\_\_

9. What would you consider your political affiliation? \_\_\_\_\_

---

Are you fiscally:

1	2	3	4	5	6	7
Very conservative	Moderately conservative	Slightly conservative	Moderate	Slightly liberal	Moderately liberal	Very liberal

Are you socially:

1	2	3	4	5	6	7
Very conservative	Moderately conservative	Slightly conservative	Moderate	Slightly liberal	Moderately liberal	Very liberal

What year are you? Freshman      Sophomore      Junior      Senior

What is your major? \_\_\_\_\_

What is your minor? \_\_\_\_\_

What is your GPA? \_\_\_\_\_

## Appendix G

### Prejudice against Blacks

1. I believe that most Blacks would discriminate against Whites if they could get away with it
2. I believe that most of the negative actions of Blacks toward Whites are due to racist feelings.
3. I believe that most Blacks would harm Whites if they could get away with it.
4. I believe that most Blacks think that they are superior to Whites.
5. I have suspected Blacks of trying to destroy something created by Whites.
6. I believe that the success of a Black person is due to their color.
7. I have blamed Blacks for my problems.
8. I have made general statements about all Blacks.
9. I believe that Blacks are selfish.
10. I believe that Black people are all alike.
11. I believe that Blacks have had an advantage just because of their color.
12. I believe that it is very unlikely that a Black person could really “like” a White.
13. Although I have befriended Blacks, I have not trusted them.
14. I believe that, despite outward appearances, most Blacks are racist.
15. I believe that most Blacks would sabotage a White’s career because they do not want Whites to succeed.

**Appendix G, Cont.**

## Prejudice against Whites

1. I believe that most Whites would discriminate against Blacks if they could get away with it
2. I believe that most of the negative actions of Whites toward Blacks are due to racist feelings.
3. I believe that most Whites would harm Blacks if they could get away with it.
4. I believe that most Whites think that they are superior to Blacks.
5. I have suspected Whites of trying to destroy something created by Blacks.
6. I believe that the success of a White person is due to their color.
7. I have blamed Whites for my problems.
8. I have made general statements about all Whites.
9. I believe that Whites are selfish.
10. I believe that White people are all alike.
11. I believe that Whites have had an advantage just because of their color.
12. I believe that it is very unlikely that a White person could really “like” a Black.
13. Although I have befriended Whites, I have not trusted them.
14. I believe that, despite outward appearances, most Whites are racist.
15. I believe that most Whites would sabotage a Black’s career because they do not want Blacks to succeed.

**Appendix H**

1. I have spent time trying to find out more about my ethnic group, such as its history, traditions, and customs.
2. I am active in organizations or social groups that include mostly members of my own ethnic group.
3. I have a clear sense of my ethnic background and what it means for me.
4. I think a lot about how my life will be affected by my ethnic group membership.
5. I am happy that I am a member of the group I belong to.
6. I have a strong sense of belonging to my own ethnic group.
7. I understand pretty well what my ethnic group membership means to me.
8. In order to learn more about my ethnic background, I have often talked to other people about my ethnic group
9. I have a lot of pride in my ethnic group
10. I participate in cultural practices of my own group, such as special food, music, or customs.
11. I feel a strong attachment towards my own ethnic group.
12. I feel good about my cultural or ethnic background.

**Appendix I**

Please circle the appropriate response or fill in the blanks given.

What is your gender?

Male      Female

2. Which of the following best describes your race or ethnicity?

Circle whatever identity applies:

African or African American

Caucasian / White

Asian or Asian American

Other: \_\_\_\_\_

Hispanic

3. How old are you? \_\_\_\_\_

What is your religion?

Christian

Jewish

Buddhist

Islamic

Hindu

Agnostic

Atheist

Spiritual but not Religious

Other: \_\_\_\_\_

How religious do you consider yourself to be?

Not very

Slightly

Somewhat

Moderately

Very much

What is your socio-economic status?

Lower

Middle

Upper

Is English your native language?

Yes      No

What is your country of origin?

Name of country: \_\_\_\_\_

9. What would you consider your political affiliation? \_\_\_\_\_

\_\_\_\_\_

10. Are you fiscally:

1	2	3	4	5	6	7
Very conservative	Moderately conservative	Slightly conservative	Moderate	Slightly liberal	Moderately liberal	Very liberal

11. Are you socially:

1	2	3	4	5	6	7
Very conservative	Moderately conservative	Slightly conservative	Moderate	Slightly liberal	Moderately liberal	Very liberal

12. What is your highest level of education?

<input type="checkbox"/> Below primary school	<input type="checkbox"/> Community (junior) College
<input type="checkbox"/> Primary/elementary school	<input type="checkbox"/> University
<input type="checkbox"/> Secondary school (high school)	<input type="checkbox"/> Graduate school

13. If you pursued higher education, what was your major? \_\_\_\_\_

14. If you pursued higher education, what was your minor? \_\_\_\_\_

15. What was your H.S. GPA? \_\_\_\_\_ (Type NA if you can't remember)

16. If you pursued higher education, what was your college GPA? \_\_\_\_\_ (Type NA if you can't remember)

17. In what industry do you work? (please check one)

- |  |   |
|--|---|
| <input type="checkbox"/> Architecture & Engineering          | <input type="checkbox"/> Legal                            |
| <input type="checkbox"/> Building & Grounds Maintenance      | <input type="checkbox"/> Life, Physical, & Social Science |
| <input type="checkbox"/> Business & Financial Operations     | <input type="checkbox"/> Management                       |
| <input type="checkbox"/> Community & Social Services         | <input type="checkbox"/> Military                         |
| <input type="checkbox"/> Computer & Mathematical             | <input type="checkbox"/> Office & Administrative Support  |
| <input type="checkbox"/> Construction & Extraction           | <input type="checkbox"/> Personal Care & Service          |
| <input type="checkbox"/> Education, Training & Library       | <input type="checkbox"/> Production                       |
| <input type="checkbox"/> Farming, Fishing, & Forestry        | <input type="checkbox"/> Protective Service               |
| <input type="checkbox"/> Food Preparation & Service          | <input type="checkbox"/> Sales & Related                  |
| <input type="checkbox"/> Related                             | <input type="checkbox"/> Student                          |
| <input type="checkbox"/> Healthcare                          | <input type="checkbox"/> Transportation & Material        |
| <input type="checkbox"/> Installation, Maintenance, & Repair | Moving  |
|  | Other (please specify)                                    |



18. What is your position in your organization? \_\_\_\_\_

19. Are you a full-time employee or part time employee? \_\_\_\_\_ Full time \_\_\_\_\_ Part time

20. How many hours a week do you work? \_\_\_\_\_ hours

21. How long have you worked in your current place of employment? \_\_\_\_\_ years  
\_\_\_\_\_ month

**Table 1. Black and White Raters' Assessment of Black and White Applicants – Archival Sample**

Black Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	0.25	0.16	0.18	0.12	
Step 2	-0.11	0.10	-0.17	0.08	
Step 3	-0.03	0.09	0.07	0.09	
Step 4	-0.11	*fixed	-0.08	*fixed	

White Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	0.32	0.16	0.15	0.11	
Step 2	-0.20	0.10	-0.02	0.08	
Step 3	0.09	0.10	-0.03	0.08	
Step 4	-0.21	*fixed	-0.10	*fixed	

*Note: There are no standard errors for Step 4, since these parameters are fixed.*

**Table 2. Portrayed Candidate Attractiveness, Masculinity, and Age**

Candidate Race	Average Attractiveness	St. Dev.	Average Masculinity	St. Dev.	Average Age	St. Dev.
Black	3.79	1.27	4.84	1.12	26.26	4.54
Black	3.53	1.43	5.26	1.37	25.84	3.20
Black	4.58	1.54	4.63	1.50	25.95	3.10
White	4.32	0.95	5.21	1.18	24.06	3.06
White	4.05	1.31	5.11	1.33	22.05	2.41
White	4.16	1.54	5.21	1.18	22.74	3.16

**Table 3. T-tests for Applicant Voices**

Voice Actor	t	df	Standard Error	Mean Whites	Mean Blacks
A15	5.92	14	0.34	4.00	2.00
A16	4.46	13	0.38	3.86	2.14
A17	6.24	14	0.37	4.33	2.00
A18	6.40	13	0.40	4.21	1.64
A19	0.34	14	0.39	3.47	3.33
A20	3.32	13	0.37	3.71	2.50

**Table 4. Depiction of Experimental Conditions**

Survey #	A15	A16	A17	A18	A19	A20
1	P1 (W)	H1 (B)	M1 (B)	H2 (W)	P2 (B)	M2 (W)
2	H1 (B)	M1 (B)	H2 (W)	P2 (B)	M2 (W)	P1 (W)
3	M1 (B)	H2 (W)	P2 (B)	M2 (W)	P1 (W)	H1 (B)
4	H2 (W)	P2 (B)	M2 (W)	P1 (W)	H1 (B)	M1 (B)
5	P2 (B)	M2 (W)	P1 (W)	H1 (B)	M1 (B)	H2 (W)
6	M2 (W)	P1 (W)	H1 (B)	M1 (B)	H2 (W)	P2 (B)
7	P1 (B)	H1 (W)	M1 (W)	H2 (B)	P2 (W)	M2 (B)
8	H1 (W)	M1 (W)	H2 (B)	P2 (W)	M2 (B)	P1 (B)
9	M1 (W)	H2 (B)	P2 (W)	M2 (B)	P1 (B)	H1 (W)
10	H2 (B)	P2 (W)	M2 (B)	P1 (B)	H1 (W)	M1 (W)
11	P2 (W)	M2 (B)	P1 (B)	H1 (W)	M1 (W)	H2 (B)
12	M2 (B)	P1 (B)	H1 (W)	M1 (W)	H2 (B)	P2 (W)

*Note: A15-A20 are codes used to denote the six voice actors. G1-M2 are codes representing manipulated interview quality, where P means “Poor”, M means “Middle”, and H means “High”. The 1’s and 2’s associated with these codes represent whether the first or second script of each quality is presented. Finally, the letter in parantheses denotes the race of the applicant. (W) represents “White”, and (B) represents “Black”.*

**Table 5. Random Coefficient Main Effects Models of Confounds Related to the Candidates**

Variable	Unstandardized Beta Weight	Standard Error	df	t
Vocal Profile – White	0.27	0.08	1178	3.46*
Vocal Profile – Black	-0.11	0.05	1178	-2.38*
Attractiveness	0.29	0.08	1178	3.87*
Masculinity	-0.72	0.10	1178	-7.08*
Order – First Half v Second Half	0.07	0.02	1174	3.20*
Order – Candidates 1 and 2 vs. Candidate 3	-0.11	0.05	1174	-2.52*
Order – Candidate 1 vs. Candidate 2	0.07	0.04	1174	1.86
Order – Candidate 4 vs. Candidates 5 and 6	-0.02	0.05	1174	-0.44
Order – Candidate 5 vs. Candidate 6	-0.04	0.04	1174	-1.05

*Note: \* indicates  $p < 0.05$*

**Table 6. Random Coefficient Interaction Models of Confounds Related to the Candidates**

Variable	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: "White" Voice</i>				
Vocal Profile – White	0.27	0.08	1176	3.46*
Candidate Race	0.83	0.31	1176	2.70*
Interaction	-0.22	0.08	1176	-2.85*
<i>Model: "Black" Voice</i>				
Vocal Profile – Black	-0.11	0.05	1176	-2.40*
Candidate Race	-0.27	0.11	1176	-2.50*
Interaction	0.10	0.05	1176	2.13*
<i>Model: Attractiveness</i>				
Attractiveness	0.16	0.16	1176	0.99
Candidate Race	0.93	0.68	1176	1.38
Interaction	-0.24	0.16	1176	-1.50
<i>Model: Masculinity</i>				
Masculinity	-0.98	0.34	1176	-2.84*
Candidate Race	0.90	1.77	1176	0.51
Interaction	-0.17	0.34	1176	-0.48
<i>Model: Order</i>				
Order – First Half v Second Half	0.07	0.02	1168	3.24*
Order – Candidates 1 and 2 vs. Candidate 3	-0.11	0.05	1168	-2.45*
Order – Candidate 2 vs. Candidate 3	0.07	0.04	1168	1.80
Order – Candidate 4 vs. Candidates 5 and 6	-0.02	0.05	1168	-0.44
Order – Candidate 5 vs. Candidate 6	-0.04	0.04	1168	-1.10
Candidate Race	-0.04	0.02	1168	-1.97*
Interaction (1,2,3 vs. 4,5,6)	-0.06	0.02	1168	-2.63*
Interaction (1,2 vs. 3)	0.10	0.05	1168	2.11*
Interaction (1 vs. 2)	0.08	0.04	1168	1.98*
Interaction (4 vs. 5,6)	0.05	0.05	1168	1.12
Interaction (5 vs. 6)	0.03	0.04	1168	0.64

Note: \* indicates  $p < 0.05$

**Table 7. Maximum Likelihood Rotated Factor Solution - White Referent**

Item Number and Text	Loading on Factor 1	Loading on Factor 2
1. Blacks are responsible for creating the racial tension that exists in the United States	0.39	0.16
4. Over the past few years, blacks have gotten more economically than they deserve.	0.81	0.18
5. Over the past few years, the government and news media have shown more respect to blacks than they deserve.	0.75	0.20
6. Blacks should not push themselves where they're not wanted.	0.30	0.32
8. I think that black people look more similar to each other than white people do.	0.10	0.70
10. It is likely that blacks will bring violence to neighborhoods when they move in.	0.18	0.43
12. Some blacks are so touchy about race that it is difficult to get along with them.	0.14	0.42



**Table 8. Maximum Likelihood Rotated Factor Solution - Black Referent**

Item Number and Text	Loading on Factor 1	Loading on Factor 2
1. Blacks are responsible for creating the racial tension that exists in the United States	0.38	0.49
4. Over the past few years, blacks have gotten more economically than they deserve.	0.59	0.30
5. Over the past few years, the government and news media have shown more respect to blacks than they deserve.	0.73	0.09
6. Blacks should not push themselves where they're not wanted.	0.41	0.26
8. I think that black people look more similar to each other than white people do.	0.06	0.45
10. It is likely that blacks will bring violence to neighborhoods when they move in.	0.29	0.48
12. Some blacks are so touchy about race that it is difficult to get along with them.	0.18	0.50

**Table 9. Multi-Group Confirmatory Factor Analyses Assessing Construct Equivalence for Prejudice Scales**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	37.24	26	0.98	0.04 (0.00 - 0.07)	0.04
Three loadings constrained	39.68	27	0.97	0.05 (0.00 - 0.07)	0.05
Four loadings constrained	41.06	28	0.97	0.05 (0.00 - 0.07)	0.05
Five loadings constrained	47.19*	29	0.96	0.05 (0.02 - 0.08)	0.05
All loadings constrained	61.90*	31	0.94	0.07 (0.04 - 0.09)	0.07

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as "1" onto each of the two factors.*

**Table 10. Multi-Group Confirmatory Factor Analyses Assessing Construct Equivalence for Motivation to Hide Prejudice Scales**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	305.02*	4	0.95	0.21 (0.19-0.23)	0.04
Two loadings constrained	305.76*	5	0.95	0.18 (0.17-0.20)	0.04
Three loadings constrained	316.23*	6	0.95	0.17 (0.15-0.19)	0.05
All loadings constrained	326.33*	7	0.95	0.16 (0.15-0.18)	0.05

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as "1" onto each of the two factors.*

**Table 11. Random Coefficient Main Effects Models of Confounds Related to the Participants**

Variable	Unstandardized Beta Weight	Standard Error	df	t
Gender	-0.02	0.03	229	-0.65
Age	0.01	0.02	207	0.47
GPA	0.09	0.06	178	1.41
Psychology Major vs. Not Psychology Major	-0.01	0.02	234	-0.41
Religion: Christian vs. Non-Christian	-0.02	0.02	233	-0.67
Religion : Jewish vs. Non-Jewish	0.05	0.03	233	1.73
Liberal vs. Not Liberal	0.00	0.02	233	0.17
Year: Freshmen & Sophomores vs. Juniors & Seniors	0.01	0.02	230	0.30
Year: Freshmen vs. Sophomores	-0.02	0.03	230	-0.68
Year: Juniors vs. Seniors	0.00	0.04	230	0.07
SES: Lower and Middle vs. Upper	0.07	0.04	231	1.76
SES: Lower vs. Middle	-0.12	0.06	231	-1.96

*Note: \* indicates  $p < 0.05$*

**Table 12. Random Coefficient Interaction Models of Confounds Related to the Participants**

Variable	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: Gender</i>				
Gender	-0.02	0.03	229	-0.65
Candidate Race	-0.04	0.03	1152	-1.42
Interaction	-0.03	0.03	1152	-1.16
<i>Model: Age</i>				
Age	0.01	0.02	207	0.47
Candidate Race	0.29	0.34	1042	0.83
Interaction	-0.02	0.02	1042	-1.00
<i>Model: GPA</i>				
GPA	0.09	0.06	178	1.42
Candidate Race	0.17	0.21	897	0.79
Interaction	-0.07	0.06	897	-1.05
<i>Model: Major</i>				
Psychology vs. Not Psychology	-0.01	0.02	234	-0.41
Candidate Race	-0.05	0.02	1177	-2.02*
Interaction	-0.01	0.02	1177	-0.56
<i>Model: Religion (Christian vs. Not)</i>				
Christian vs. Non-Christian	-0.02	0.02	233	-0.67
Candidate Race	-0.05	0.02	1172	-2.08*
Interaction	0.03	0.02	1172	1.16
<i>Model: Religion (Jewish vs. Not)</i>				
Jewish vs. Non-Jewish	0.05	0.03	233	1.72
Candidate Race	-0.05	0.03	1172	-1.97*
Interaction	-0.02	0.03	1172	-0.58
<i>Model: Liberal vs. Not Liberal</i>				
Liberal vs. Not Liberal	0.00	0.02	233	0.17
Candidate Race	-0.05	0.02	1172	-2.02*
Interaction	0,01	0,02	1172	0.25

Note: \* indicates  $p < 0.05$

**Table 12, Cont. Random Coefficient Interaction Models of Confounds Related to the Participants**

Confound	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: Year in College</i>				
Year: Freshmen & Sophomores vs. Juniors & Seniors	0.01	0.02	230	0.30
Year: Freshmen vs. Sophomores	-0.02	0.03	230	-0.68
Year: Juniors vs. Seniors	0.00	0.04	230	0.07
Candidate Race	-0.06	0.02	1165	-2.40*
Interaction (1&2 vs. 3&4)	0.04	0.02	1165	1.82
Interaction (1 vs. 2)	0.01	0.03	1165	0.85
Interaction (3 vs. 4)	0.02	0.04	1165	0.50
<i>Model: SES</i>				
SES: Lower and Middle vs. Upper	0.07	0.04	231	1.76
SES: Lower vs. Middle	-0.12	0.06	231	-1.96
Candidate Race	-0.03	0.04	1166	-0.66
Interaction (Lower & Middle vs. Upper)	-0.00	0.04	1166	-0.04
Interaction (Lower vs. Middle)	0.03	0.06	1166	0.43

Note: \* indicates  $p < 0.05$

**Table 13. Correlations amongst Study Variables**

	Mean	Std.	1	2	3	4	5	6	7	8	9
1. Participant Race	0.42	0.91	-								
2. Average Rating – White Candidates	3.53	0.40	0.03	(0.68)							
3. Average Rating – Black Candidates	3.62	0.40	0.09	0.29*	(0.66)						
4. Absolute difficulty parameters – White Candidates	1.19	1.25	-0.01	0.20*	0.07	-					
5. Absolute difficulty parameters – Black Candidates	1.40	1.44	0.04	0.11	0.17*	-0.02	-				
4. Prejudice against Whites (Overall)	3.05	0.93	-0.36*	-0.01	-0.08	-0.06	0.08	(0.69)			
5. Prejudice against Blacks (Overall)	2.72	0.91	0.24*	0.05	0.02	-0.09	0.10	0.24*	(0.70)		
6. Motivation to Hide Prejudice against Whites	3.47	1.49	-0.04	0.17*	-0.02	-0.01	-0.07	0.17*	0.15*	(0.86)	
7. Motivation to Hide Prejudice against Blacks	3.83	1.37	0.26*	0.19*	0.03	0.06	0.02	0.05	0.38*	0.67*	(0.81)

Note: \* indicates  $p < 0.05$ ; Race: -1 = Black, 1 = White. Reliabilities are along the diagonal,  $n = 232$

**Table 14. Black and White Raters' Assessment of Black and White Applicants – Student Sample**

Black Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	-1.70	0.10	-1.33	0.10	
Step 2	0.05	0.09	-0.55	0.09	
Step 3	0.26	0.09	0.77	0.10	
Step 4	1.40	*fixed	1.11	*fixed	

White Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	-1.40	0.06	-1.73	0.07	
Step 2	-0.25	0.06	-0.22	0.06	
Step 3	0.40	0.06	0.63	0.06	
Step 4	1.24	*fixed	1.33	*fixed	



**Table 15. Predictors of Difficulty Parameters for Latent Class 1 (DTF against blacks, not in favor of whites)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	2.94	0.11	25.91*
Second Effects-Coded Variable	-3.24	0.23	-14.41*
Third Effects-Coded Variable	-0.41	0.28	-1.46
Candidate Race	-0.08	0.02	-3.40*
Interaction between the first effects-coded variable and candidate race	-3.52	0.11	-32.27*
Interaction between the second effects-coded variable and candidate race	3.06	0.19	16.33*
Interaction between the third effects-coded variable and candidate race	0.57	0.22	2.53*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 16. Predictors of Difficulty Parameters for Latent Class 2 (DTF against whites, somewhat in favor of blacks)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	6.07	0.24	25.40*
Second Effects-Coded Variable	0.04	0.67	0.07
Third Effects-Coded Variable	-4.73	0.29	-16.44*
Candidate Race	0.29	0.08	3.53*
Interaction between the first effects-coded variable and candidate race	0.99	0.21	4.79*
Interaction between the second effects-coded variable and candidate race	3.45	0.36	9.61*
Interaction between the third effects-coded variable and candidate race	-4.10	0.14	-28.60*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 17. Predictors of Difficulty Parameters for Latent Class 3 (No DTF)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	-0.32	0.06	-5.56*
Second Effects-Coded Variable	-0.39	0.05	-7.29*
Third Effects-Coded Variable	0.13	0.05	2.41*
Candidate Race	-0.01	0.01	-0.46
Interaction between the first effects-coded variable and candidate race	-0.02	0.05	-0.37
Interaction between the second effects-coded variable and candidate race	-0.07	0.05	-1.51
Interaction between the third effects-coded variable and candidate race	0.15	0.06	2.55*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 18. Predictors of Difficulty Parameters for Latent Class 4 (DTF against blacks, not in favor of whites)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	-0.22	0.17	-1.30
Second Effects-Coded Variable	3.68	0.32	11.38*
Third Effects-Coded Variable	-3.88	0.25	-15.25*
Candidate Race	0.04	0.04	0.86
Interaction between the first effects-coded variable and candidate race	0.14	0.14	0.99
Interaction between the second effects-coded variable and candidate race	-3.22	0.30	-10.83*
Interaction between the third effects-coded variable and candidate race	3.20	0.28	11.40*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 19. Predictors of Difficulty Parameters for Latent Class 5 (DTF against whites, not in favor of blacks)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	3.11	0.40	7.87*
Second Effects-Coded Variable	-3.72	0.35	-10.62*
Third Effects-Coded Variable	0.22	0.18	1.21
Candidate Race	-0.02	0.06	-0.24
Interaction between the first effects-coded variable and candidate race	3.08	0.40	7.62*
Interaction between the second effects-coded variable and candidate race	-3.35	0.30	-11.05*
Interaction between the third effects-coded variable and candidate race	0.13	0.12	1.04

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 20. Confirmatory Factor Analyses for Prejudice Scales, Adult Replication**

	$\chi^2$	df	CFI	RMSEA	SRMR
<b>Black Prejudice Models</b>					
One Overall Factor	46.01*	14	0.91	0.11 (0.08 – 0.15)	0.05
Two Separate Factors	40.75*	13	0.92	0.11 (0.07 – 0.15)	0.05
Three Separate Factors	28.93*	11	0.95	0.10 (0.05 – 0.14)	0.05
<b>White Prejudice Models</b>					
One Overall Factor	52.51*	14	0.87	0.12 (0.09 – 0.16)	0.07
Two Separate Factors	38.19*	13	0.91	0.10 (0.07 – 0.14)	0.06
Three Separate Factors	32.27*	11	0.93	0.10 (0.06 – 0.15)	0.05

*Note: \* indicates that the  $\chi^2$  is significant at 0.05*

**Table 21. Multi-Group Confirmatory Factor Analyses Assessing Construct Equivalence for Prejudice Scales, Adult Replication**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	78.94*	26	0.92	0.11 (0.08 – 0.13)	0.06
Three loadings constrained	79.22*	27	0.92	0.10 (0.08 – 0.13)	0.06
Four loadings constrained	79.61*	28	0.92	0.10 (0.08 – 0.13)	0.06
Five loadings constrained	82.55*	29	0.91	0.10 (0.08 – 0.13)	0.06
Six loadings constrained	88.90*	30	0.91	0.10 (0.08 – 0.13)	0.07
All loadings constrained	90.57*	31	0.90	0.10 (0.08 – 0.13)	0.07

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as “1” onto each of the two factors.*

**Table 22. Maximum Likelihood One-Factor Solution, 15 Items**

Item	Loading for Black Referent	Loading for White Referent
Item 1: I believe that most Blacks/Whites would discriminate against Whites/Blacks if they could get away with it	0.61	0.73
Item 2: I believe that most of the negative actions of Blacks/Whites toward Whites/Blacks are due to racist feelings.	0.33	0.59
Item 3: I believe that most Blacks/Whites would harm Whites/Blacks if they could get away with it.	0.72	0.77
Item 4: I believe that most Blacks/Whites think that they are superior to Whites/Blacks.	0.79	0.71
Item 5: I have suspected Blacks/Whites of trying to destroy something created by Whites/Blacks.	0.70	0.79
Item 6: I believe that the success of a Black/White person is due to their color.	0.55	0.65
Item 7: I have blamed Blacks/Whites for my problems.	0.68	0.63
Item 8: I have made general statements about all Blacks/Whites.	0.44	0.61
Item 9: I believe that Blacks/Whites are selfish.	0.73	0.64
Item 10: I believe that Black/White people are all alike.	0.80	0.73
Item 11: I believe that Blacks/Whites have had an advantage just because of their color.	0.62	0.56
Item 12: I believe that it is very unlikely that a Black/White person could really "like" a White/Black.	0.78	0.68
Item 13: Although I have befriended Blacks/Whites, I have not trusted them.	0.82	0.73
Item 14: I believe that, despite outward appearances, most Blacks/Whites are racist.	0.82	0.81
Item 15: I believe that most Blacks/Whites would sabotage a White's/Black's career because they do not want Whites/Black to succeed.	0.82	0.91



**Table 23. Confirmatory Factor Analyses for Prejudice Scales, Adult Extension**

	$\chi^2$	df	CFI	RMSEA	SRMR
<b>Black Prejudice Models</b>					
One Overall Items	371.30*	90	0.83	0.13 (0.12 – 0.15)	0.07
One Factor Parcels	12.41*	5	0.99	0.09 (0.03 – 0.15)	0.02
<b>White Prejudice Models</b>					
One Overall Items	574.43*	90	0.74	0.17 (0.16 – 0.19)	0.10
One Factor Parcels	2.05	5	1.00	0.00 (0.00 – 0.06)	0.01

*Note: \* indicates that the  $\chi^2$  is significant at 0.05*

**Table 24. Multi-Group Confirmatory Factor Analyses Assessing Construct Equivalence for Prejudice Scales, Adult Extension**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	14.46	10	1.00	0.05 (0.00 – 0.10)	0.01
Restricting the third parcel	16.12	11	1.00	0.05 (0.00 – 0.10)	0.02
Restricting the fourth parcel	19.12	12	1.00	0.06 (0.00 – 0.10)	0.03

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as “1” onto each of the two factors.*

**Table 25: External Motivation to Hide Prejudice, Adult Sample**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	18.85*	4	0.98	0.14 (0.08 – 0.21)	0.04
Two loadings constrained	19.46*	5	0.98	0.13 (0.07 – 0.19)	0.04
Three loadings constrained	21.03*	6	0.98	0.12 (0.07 – 0.17)	0.05
All loadings constrained	28.49*	7	0.97	0.13 (0.08 – 0.18)	0.06

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as “1” onto each of the two factors.*

**Table 26. Ethnic Identity CFA**

	$\chi^2$	df	CFI	RMSEA	SRMR
Overall Model					
One Factor – Items	275.72*	54	0.83	0.15 (0.13 – 0.17)	0.08
One Factor – Parcels	0.16	2	1.00	0.00 (0.00 – 0.05)	0.00

*Note: \* indicates that the  $\chi^2$  is significant at 0.05*

**Table 27. Multi-Group Confirmatory Factor Analyses Assessing Construct Equivalence for Ethnic Identity**

	$\chi^2$	df	CFI	RMSEA	SRMR
Models					
No constraints	4.02	4	1.00	0.01 (0.00 – 0.16)	0.01
Second parcel constrained loading	5.69	5	1.00	0.04 (0.00 – 0.16)	0.05
Third parcel constrained loading	9.33	6	0.99	0.08 (0.00 – 0.17)	0.09
Fourth parcel constrained loading	10.83	7	0.99	0.08 (0.00 – 0.16)	0.09
Intercepts constrained	18.32*	10	0.98	0.10 (0.00 – 0.16)	0.12
Residuals constrained	21.55	13	0.98	0.09 (0.00 – 0.15)	0.13
Factor variance constrained	22.74	14	0.98	0.08 (0.00 – 0.14)	0.16

*Note: \* indicates that the  $\chi^2$  is significant at 0.05. Number of loadings constrained includes the two loadings that were constrained to load as “1” onto each of the two factors.*

**Table 28. Random Coefficient Main Effects Models of Confounds Related to the Participants**

Variable	Unstandardized Beta Weight	Standard Error	df	t
Gender	0.06	0.03	182	1.83
Age	-0.00	0.00	181	-0.01
High School GPA	0.00	0.00	100	1.04
College GPA	-0.00	0.10	94	-0.04
Business Major vs. Not Business Major	0.01	0.05	138	0.19
Religion: Christian vs. Non-Christian	0.04	0.03	182	1.43
Liberal vs. Not Liberal	-0.00	0.03	182	-0.02
SES: Lower and Middle vs. Upper	0.06	0.12	181	0.48
SES: Lower vs. Middle	-0.05	0.03	181	-1.34
Education: High School and Below vs. Post-High School	-0.48	0.18	179	-2.70*
Education: Elementary School vs. Secondary Education	-0.46	0.14	179	-3.17*
Education: Community College vs. University and Above	0.08	0.05	179	1.59
Education: Undergraduate vs. Graduate	-0.05	0.04	179	-1.16
English as a native language vs. Other native language	-0.05	0.06	182	-0.77
USA as country of origin vs. Other country of origin	-0.06	0.05	181	-1.29
Full time vs. Not full time	0.01	0.04	168	0.16
Hours worked per week	0.01	0.00	132	1.71
Prior experience interviewing	-0.01	0.04	181	-0.17
Prior experience as an EMT, Paramedic, or Firefighter	-0.09	0.07	182	-1.22

*Note: \* indicates  $p < 0.05$*

**Table 29. Random Coefficient Interaction Models of Confounds Related to the Participants**

Variable	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: Gender</i>				
Gender	0.06	0.03	182	1.83
Candidate Race	0.02	0.02	918	0.68
Interaction	-0.02	0.02	918	-0.88
<i>Model: Age</i>				
Age	-0.00	0.00	181	-0.01
Candidate Race	-0.09	0.08	913	-1.12
Interaction	0.00	0.00	913	1.28
<i>Model: High School GPA</i>				
High School GPA	0.00	0.00	100	1.04
Candidate Race	0.01	0.03	508	0.37
Interaction	0.00	0.00	508	0.51
<i>Model: College GPA</i>				
College GPA	-0.00	0.10	94	-0.04
Candidate Race	0.23	0.29	478	0.81
Interaction	-0.07	0.08	478	-0.83
<i>Model: Major</i>				
Business vs. Not Business	0.01	0.05	138	0.19
Candidate Race	0.06	0.03	698	1.80
Interaction	0.05	0.03	698	1.55
<i>Model: Religion (Christian vs. Not)</i>				
Christian vs. Non-Christian	0.04	0.03	182	1.43
Candidate Race	0.01	0.02	918	0.62
Interaction	-0.02	0.02	918	-0.68
<i>Model: Democrat vs. Not Democrat</i>				
Liberal vs. Not Liberal	-0.00	0.03	182	-0.02
Candidate Race	0.01	0.02	918	0.66
Interaction	0.04	0.02	918	1.79

Note: \* indicates  $p < 0.05$

**Table 29, Cont. Random Coefficient Interaction Models of Confounds Related to the Participants**

Confound	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: SES</i>				
SES: Lower and Middle vs. Upper	0.06	0.12	181	0.48
SES: Lower vs. Middle	-0.05	0.03	181	-1.34
Candidate Race	-0.26	0.09	917	-2.94*
Interaction (Lower & Middle vs. Upper)	0.26	0.09	917	2.98*
Interaction (Lower vs. Middle)	-0.03	0.03	917	-1.18
<i>Model: Highest Education</i>				
Pre-Secondary vs. Post-Secondary	-0.48	0.18	179	-2.70*
Primary vs. Secondary	-0.46	0.14	179	-3.17*
Community vs. Four-year and beyond	0.08	0.05	179	1.59
Undergraduate vs. Graduate	-0.05	0.04	179	-1.16
Candidate Race	0.06	0.05	915	1.26
Interaction (Pre vs. Post-Secondary)	0.13	0.14	915	0.97
Interaction (Primary vs. Secondary)	0.15	0.11	915	1.39
Interaction (Community vs. at least four years)	-0.01	0.04	915	-0.34
Interaction (Undergraduate vs. Graduate)	-0.00	0.03	915	-0.11
<i>Model: English as a native language</i>				
English as a native language	-0.05	0.06	182	-0.77
Candidate Race	0.06	0.04	918	1.26
Interaction	-0.05	0.04	918	-1.21
<i>Model: USA as country of origin</i>				
USA as country of origin	-0.06	0.05	181	-1.29
Candidate Race	0.03	0.04	913	0.84
Interaction	-0.03	0.04	913	-0.72

Note: \* indicates  $p < 0.05$



**Table 29, Cont. Random Coefficient Interaction Models of Confounds Related to the Participants**

Confound	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: Full time status</i>				
Full time vs. Not full time	0.01	0.04	168	0.16
Candidate Race	0.01	0.03	848	0.20
Interaction (Lower vs. Middle)	0.01	0.03	848	0.36
<i>Model: Hours worked per week</i>				
Hours worked per week	0.01	0.00	132	1.71
Candidate Race	0.03	0.08	668	0.39
Interaction (Pre vs. Post-Secondary)	-0.00	0.00	668	-0.28
<i>Model: Prior interviewing experience</i>				
Experience interviewing candidates	-0.01	0.04	181	-0.17
Candidate Race	0.03	0.03	913	1.12
Interaction	0.04	0.03	913	1.48
<i>Model: Prior job experience</i>				
Experience as EMT, paramedic, or firefighter	-0.09	0.07	182	-1.22
Candidate Race	0.06	0.05	918	1.07
Interaction	0.05	0.05	918	0.97

Note: \* indicates  $p < 0.05$

**Table 30. Between-Participant Correlations, Adult Sample**

	Mean	St Dev.	1	2	3	4	5	6	7	8	9	10	11	12
1. Participant Race	0.34	0.94	-											
2. Average Rating - White Candidates	3.62	0.46	-0.08	(0.60)										
3. Average Rating - Black Candidates	3.60	0.48	0.07	0.49*	(0.62)									
4. Absolute Difference Scores - White Candidates	1.39	1.47	0.03	0.01	0.08	-								
5. Absolute Difference Scores - Black Candidate	1.19	1.38	0.06	0.15	0.03	0.06	-							
6. Prejudice against Whites - Replicated from Study 2	3.19	1.01	-0.32*	-0.04	-0.14	-0.11	-0.18*	(0.71)						
7. Prejudice against Blacks - Replicated from Study 2	2.74	1.03	0.14	-0.03	0.08	0.05	-0.14	0.17*	(0.79)					
8. Prejudice against Whites - New Scale	2.98	1.36	-0.40*	-0.06	-0.23*	-0.06	-0.21*	0.59*	0.02	(0.88)				
9. Prejudice against Blacks - New scale	2.09	1.03	-0.06	-0.10	-0.07	-0.10	-0.12	0.30*	0.73*	0.28*	(0.88)			
10. External motivation to hide prejudice against Whites	2.99	1.46	-0.21*	0.01	-0.17*	-0.04	-0.14	0.30*	0.25*	0.24*	0.32*	(0.86)		
11. External motivation to hide prejudice against Blacks	3.17	1.39	0.00	-0.09	-0.20*	-0.01	-0.06	0.31*	0.32*	0.19*	0.41*	0.62*	(0.77)	
12. Ethnic Identity	4.28	1.59	-0.28*	0.09	0.10	0.05	0.04	0.20*	0.13	0.18*	0.21*	0.11	0.05	(0.91)

*Note: \* indicates significant at 0.05, n = 181*

**Table 31. Black and White Raters' Assessment of Black and White Applicants – Adult Sample**

Black Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	-1.71	0.11	-1.35	0.10	
Step 2	-0.64	0.09	-0.76	0.09	
Step 3	0.87	0.10	0.65	0.09	
Step 4	1.48	*fixed	1.47	*fixed	

White Raters					
	Black Applicants		White Applicants		
	Difficulty Parameter	Standard Error	Difficulty Parameter	Standard Error	
Step 1	-1.72	0.08	-1.56	0.08	
Step 2	-0.51	0.07	-0.50	0.07	
Step 3	0.65	0.07	0.70	0.07	
Step 4	1.58	*fixed	1.36	*fixed	

*Note: There are no standard errors for Step 4, since these parameters are fixed.*

**Table 32. Predictors of Difficulty Parameters for Latent Class 1 (DTF against blacks, DTF against whites)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	2.50	0.11	23.14*
Second Effects-Coded Variable	2.11	0.37	5.65*
Third Effects-Coded Variable	-4.33	0.19	-22.51*
Candidate Race	-0.64	0.11	-5.92*
Interaction between the first effects-coded variable and candidate race	-2.39	0.11	-22.08*
Interaction between the second effects-coded variable and candidate race	4.62	0.37	12.65*
Interaction between the third effects-coded variable and candidate race	-2.88	0.17	-17.12*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 33. Predictors of Difficulty Parameters for Latent Class 2 (DTF against whites, not in favor of blacks)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	4.39	0.53	8.34*
Second Effects-Coded Variable	-4.37	0.53	-8.32*
Third Effects-Coded Variable	-0.44	0.21	-2.09*
Candidate Race	-0.16	0.08	-2.01*
Interaction between the first effects-coded variable and candidate race	2.07	0.54	3.85*
Interaction between the second effects-coded variable and candidate race	-2.31	0.44	-5.21*
Interaction between the third effects-coded variable and candidate race	-0.07	0.25	-0.27

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 34. Predictors of Difficulty Parameters for Latent Class 3 (No DTF)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	-0.33	0.07	-4.70*
Second Effects-Coded Variable	-0.39	0.07	-5.90*
Third Effects-Coded Variable	0.29	0.07	4.54*
Candidate Race	-0.01	0.02	-0.34
Interaction between the first effects-coded variable and candidate race	-0.01	0.06	-0.19
Interaction between the second effects-coded variable and candidate race	0.07	0.06	1.21
Interaction between the third effects-coded variable and candidate race	0.02	0.07	0.27

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 35. Predictors of Difficulty Parameters for Latent Class 4 (DTF against whites, not in favor of blacks)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	3.11	0.09	35.94*
Second Effects-Coded Variable	-3.82	0.20	-19.14*
Third Effects-Coded Variable	0.26	0.22	1.20
Candidate Race	-0.02	0.02	-0.70
Interaction between the first effects-coded variable and candidate race	-3.19	0.10	-33.28*
Interaction between the second effects-coded variable and candidate race	3.07	0.11	28.29*
Interaction between the third effects-coded variable and candidate race	0.24	0.19	1.23

Note: \* Indicates that the t-value is statistically significant at the 0.05 level

**Table 36. Predictors of Difficulty Parameters for Latent Class 5 (DTF against whites, not in favor of blacks)**

Predictor	Unstandardized Beta weight	Standard Error	t
<i>Within</i>			
First Effects-Coded Variable	0.66	0.72	0.92
Second Effects-Coded Variable	3.27	1.02	3.22*
Third Effects-Coded Variable	-4.30	0.40	-10.86*
Candidate Race	0.17	0.11	1.49
Interaction between the first effects-coded variable and candidate race	0.84	1.00	0.83
Interaction between the second effects-coded variable and candidate race	-3.43	1.06	-3.24*
Interaction between the third effects-coded variable and candidate race	2.90	0.41	7.13*

Note: \* Indicates that the t-value is statistically significant at the 0.05 level



**Table 37. Estimated Marginal Means of Prejudice and Motivation to Hide Prejudice across Classes**

Variable	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4	Latent Class 5
<b>Replication</b>					
Prejudice against Whites	2.59	3.41	3.97	3.15	2.84
Prejudice against Blacks	2.62	2.22	3.00	2.63	3.24
Motivation to hide Prejudice against Whites	2.50	3.34	3.20	3.11	2.80
Motivation to hide Prejudice against Blacks	3.39	3.04	2.95	3.45	3.02
Ethnic Identity	4.66	4.77	3.87	3.99	4.64
<b>Extension</b>					
Prejudice against Whites	2.55	2.89	3.36	3.04	3.07
Prejudice against Blacks	2.49	1.58	2.63	1.68	2.08
Motivation to hide Prejudice against Whites	2.43	3.42	3.26	3.09	2.76
Motivation to hide Prejudice against Blacks	3.31	3.10	2.88	3.51	3.05
Ethnic Identity	4.47	4.64	3.96	4.06	4.80

**Table 38. Estimated Marginal Means of Prejudice and Motivation to Hide Prejudice across Samples**

Variable	Student Sample	Adult Sample – Replication	Adult Sample - Extension
<b>No DTF</b>			
Prejudice against Whites – Replicated	3.18	3.97	
Prejudice against Blacks – Replicated	2.68	3.00	
Prejudice against Whites – Extended			3.36
Prejudice against Blacks – Extended			2.63
Motivation to hide Prejudice against Whites	3.64	3.20	3.26
Motivation to hide Prejudice against Blacks	3.56	2.95	2.88
Ethnic Identity		3.87	3.96
<b>DTF against Whites, not in favor of Blacks</b>			
Prejudice against Whites – Replicated	3.10	3.41	
Prejudice against Blacks – Replicated	2.69	2.22	
Prejudice against Whites – Extended			2.89
Prejudice against Blacks – Extended			1.58
Motivation to hide Prejudice against Whites	3.43	3.34	3.42
Motivation to hide Prejudice against Blacks	3.49	3.04	3.10
Ethnic Identity		4.77	4.64
<b>DTF against Whites, DTF against Blacks</b>			
Prejudice against Whites – Replicated	2.55	2.59	
Prejudice against Blacks – Replicated	2.28	2.62	
Prejudice against Whites – Extended			2.55
Prejudice against Blacks – Extended			2.49
Motivation to hide Prejudice against Whites	3.54	2.50	2.43
Motivation to hide Prejudice against Blacks	4.78	3.39	3.31
Ethnic Identity		4.66	4.47

**Table 38, cont. Estimated Marginal Means of Prejudice and Motivation to Hide Prejudice across Samples**

Variable	Student Sample	Adult Sample – Replication	Adult Sample - Extension
DTF against blacks, not in favor of whites ( <i>between step 1 and 2</i> )			
Prejudice against Whites – Replicated	3.03	3.15	
Prejudice against Blacks – Replicated	3.02	2.63	
Prejudice against Whites – Extended			3.04
Prejudice against Blacks – Extended			1.68
Motivation to hide Prejudice against Whites	3.45	3.11	3.09
Motivation to hide Prejudice against Blacks	3.75	3.45	3.51
Ethnic Identity		3.99	4.06
DTF against blacks, not in favor of whites ( <i>between step 2 and 3</i> )			
Prejudice against Whites – Replicated	3.40	2.84	
Prejudice against Blacks – Replicated	2.93	3.24	
Prejudice against Whites – Extended			3.07
Prejudice against Blacks – Extended			2.08
Motivation to hide Prejudice against Whites	3.28	2.80	2.76
Motivation to hide Prejudice against Blacks	3.57	3.02	3.05
Ethnic Identity		4.64	4.80

**Table 39. Chi-Square Tests of Demographic Differences in Latent Profiles (Student)**

	Group 1	Group 2	Group 3	Group 4	Group 5	Test of significance
% Female	61.10%	80.00%	67.40%	81.00%	81.30%	$\chi^2(4) = 3.75$
% Christian	47.40%	50.00%	57.50%	47.60%	56.30%	$\chi^2(24) = 18.69$
% Jewish	15.80%	40.00%	23.40%	28.60%	25.00%	
Mean religiosity	2.74	3.00	2.75	2.86	3.06	$\chi^2(4) = 1.11$
% Middle class	84.20%	70.00%	81.00%	81.00%	87.50%	$\chi^2(8) = 7.62$
% Lower class	10.50%	0.00%	3.60%	0.00%	6.30%	
% Democrat	63.20%	70.00%	51.90%	75.00%	68.80%	$\chi^2(16) = 7.08$
% Psychology Major	52.60%	20.00%	42.33%	52.38%	43.75%	$\chi^2(4) = 3.75$
% Who guessed that race was involved	100.00%	80.00%	96.00%	90.48%	100.00%	$\chi^2(8) = 23.91^*$
% Who guessed that race was involved during the rating portion of the study	78.90%	83.30%	94.40%	90.00%	100.00%	$\chi^2(8) = 10.43,$
% Who said that guessing that race was involved affected their ratings	22.20%	20.00%	13.20%	5.90%	33.30%	$\chi^2(4) = 6.11,$
Mean age	19.56	18.83	19.35	19.40	19.00	$\chi^2(4) = 2.82,$
Mean GPA	3.50	3.18	3.31	3.39	3.21	$\chi^2(4) = 4.80,$

Note: \* $p < 0.05$

**Table 40. Chi-Square Tests of Demographic Differences in Latent Profiles (Adult)**

	Group 1	Group 2	Group 3	Group 4	Group 5	Test of significance
% Female	84.50%	77.80%	61.30%	50.00%	77.80%	$\chi^2(4) = 5.83$
% Christian	46.20%	72.20%	70.60%	50.00%	55.60%	$\chi^2(32) = 43.65$
Mean religiosity	3.08	3.00	2.91	2.30	2.89	$\chi^2(4) = 2.07$
% Middle class	69.20%	72.20%	73.10%	60.00%	100.00%	$\chi^2(8) = 10.03$
% Lower class	30.80%	22.20%	26.10%	30.00%	0.00%	
% Democrat	38.50%	27.80%	46.20%	20.00%	44.40%	$\chi^2(20) = 15.19$
% Undergraduate	23.10%	27.80%	46.20%	30.00%	44.40%	$\chi^2(16) = 29.44^*$
% 2 year college	38.50%	11.10%	24.40%	20.00%	33.30%	
% High school	30.80%	16.70%	20.20%	10.00%	11.10%	
% Experience in relevant field	0.00%	0.00%	4.20%	10.00%	0.00%	$\chi^2(4) = 2.84$
% Experience interviewing	30.80%	11.10%	22.90%	10.00%	22.20%	$\chi^2(4) = 2.74$
% Business Major	7.70%	16.70%	13.40%	0.00%	11.10%	$\chi^2(16) = 13.66$
% Work full-time	72.70%	87.50%	79.30%	75.00%	66.70%	$\chi^2(4) = 1.84$
% Who guessed that race was involved	92.30%	94.40%	79.80%	90.00%	88.90%	$\chi^2(4) = 3.94$
% Who guessed that race was involved during the rating portion of the study	84.60%	72.20%	59.70%	60.00%	77.80%	$\chi^2(4) = 4.70$
% Who said that guessing that race was involved affected their ratings	23.10%	16.70%	13.40%	10.00%	11.10%	$\chi^2(4) = 1.20$
Mean age	39.15	32.67	32.63	36.10	36.33	$\chi^2(4) = 6.13$
Mean College GPA	3.51	3.43	3.43	3.62	3.51	$\chi^2(4) = 1.63$

Note: \*p &lt; 0.05

**Table 41. Random Coefficient Main Effects Models of Confounds Related to the Design**

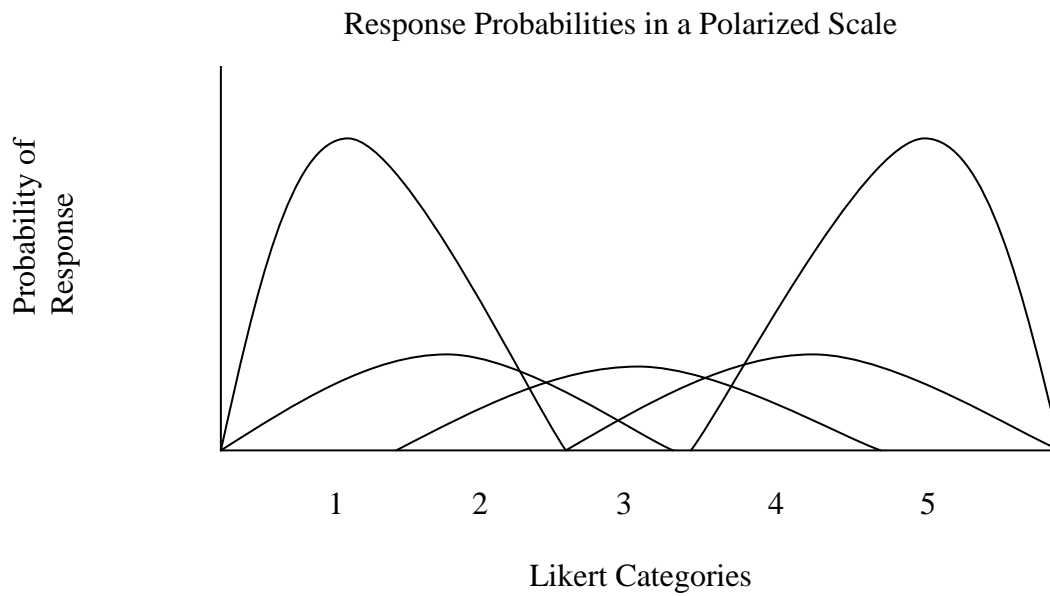
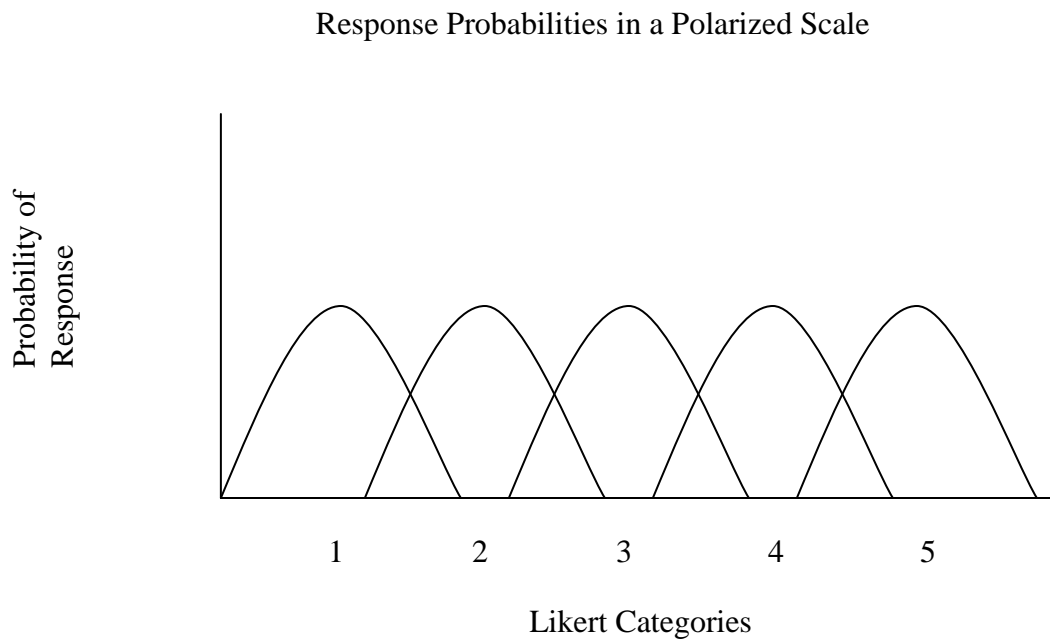
Variable	Unstandardized Beta Weight	Standard Error	df	t
Vocal Profile – White	0.21	0.08	919	2.76*
Vocal Profile – Black	-0.14	0.04	919	-3.20*
Attractiveness	0.21	0.07	919	2.83*
Masculinity	-0.46	0.10	919	-4.60*
Order – First Half v Second Half	0.06	0.02	915	2.93*
Order – Candidates 1 and 2 vs. Candidate 3	0.04	0.04	915	0.94
Order – Candidate 1 vs. Candidate 2	0.04	0.04	915	1.11
Order – Candidate 4 vs. Candidates 5 and 6	-0.10	0.04	915	-2.19
Order – Candidate 5 vs. Candidate 6	-0.02	0.04	915	-0.57

*Note: \* indicates  $p < 0.05$*

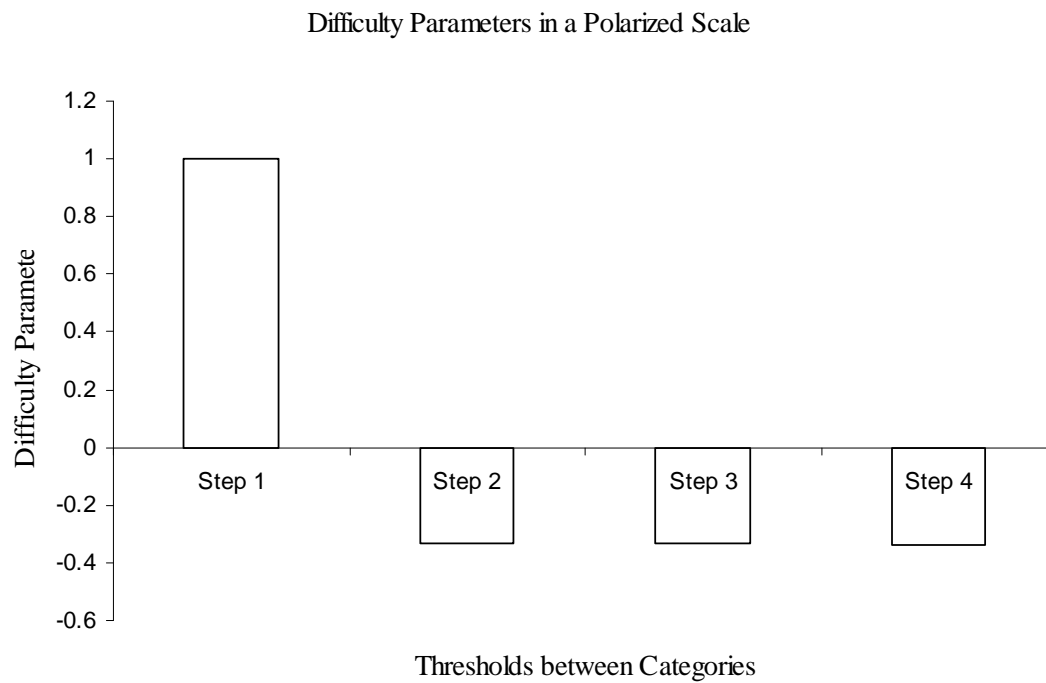
**Table 42. Random Coefficient Interaction Models of Confounds Related to the Candidates**

Variable	Unstandardized Beta Weight	Standard Error	df	t
<i>Model: "White" Voice</i>				
Vocal Profile – White	0.21	0.08	917	2.74*
Candidate Race	0.33	0.31	917	1.07
Interaction	-0.08	0.08	917	-1.04
<i>Model: "Black" Voice</i>				
Vocal Profile – Black	-0.14	0.04	917	-3.11*
Candidate Race	-0.05	0.11	917	-0.41
Interaction	0.02	0.05	917	0.50
<i>Model: Attractiveness</i>				
Attractiveness	-0.31	0.16	917	-2.02*
Candidate Race	2.53	0.65	917	3.89*
Interaction	-0.61	0.16	917	-3.91*
<i>Model: Masculinity</i>				
Masculinity	-1.87	0.33	917	-5.66*
Candidate Race	6.88	1.70	917	4.04*
Interaction	-1.32	0.33	917	-3.99*
<i>Model: Order</i>				
Order – First Half v Second Half	0.07	0.02	909	3.13*
Order – Candidates 1 and 2 vs. Candidate 3	0.04	0.04	909	0.96
Order – Candidate 2 vs. Candidate 3	0.05	0.04	909	1.33
Order – Candidate 4 vs. Candidates 5 and 6	-0.08	0.04	909	-1.87
Order – Candidate 5 vs. Candidate 6	-0.02	0.04	909	-0.51
Candidate Race	0.01	0.02	909	0.62
Interaction (1,2,3 vs. 4,5,6)	0.04	0.02	909	1.95
Interaction (1,2 vs. 3)	0.06	0.05	909	1.26
Interaction (1 vs. 2)	0.00	0.04	909	0.12
Interaction (4 vs. 5,6)	0.09	0.05	909	1.92
Interaction (5 vs. 6)	-0.01	0.04	909	-0.28

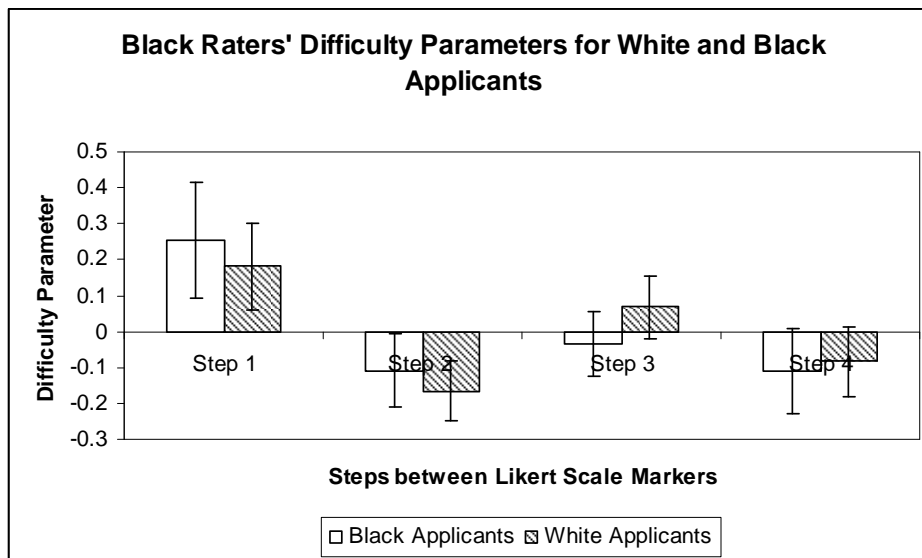
Note: \* indicates  $p < 0.05$

**Figure 1.1 Response Probabilities in a Polarized Scale****Figure 1.2 Response Probabilities Given Equivalent Usage of Scale Markers**



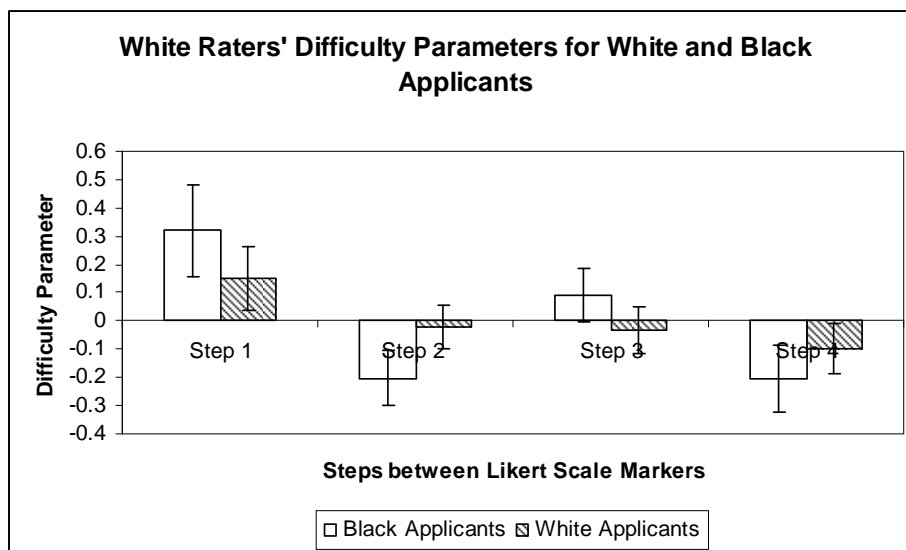
**Figure 1.3 Difficulty Parameters in Polarized Scale**

**Figure 2.1 Black Raters' Assessment of Black and White Applicants – Archival Sample**




*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*

**Figure 2.2 White Raters' Assessment of Black and White Applicants – Archival Sample**



*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*

Figure 3. Screenshot of Rating Page



UNIVERSITY OF  
MARYLAND

**Candidate 1 - Situation1**


**Male**  
**Caucasian**  
**25 Years Old**

Assume that you are an entry-level firefighter. You work on a 24 hour shift. During the shift, you and your co-workers are required to work and live closely together. After you eat dinner with your coworkers each night, everyone who did not help cook the meal is supposed to help clean up the dirty dishes and the kitchen. One of the other entry-level firefighters who works on your shift seems to always avoid cleaning anything by staying seated at the kitchen table until all of the cleaning has been done. You like this co-worker. You realize, however, that he is not doing his fair share of cleaning and he is beginning to irritate you and others.

Follow-Up Questions  
What actions, if any, would you take in this situation and why? Would you say anything to this co-worker? Why or why not? What would you say? Would you involve anyone else? If so, who and why? If not, why not?



[Click here for Candidate #1's Response to Situation #1](#)



Notes

Rate the candidate on his or her response to the first situation using the benchmark guidelines below.

1 - Unacceptable	2 - Less than Acceptable	3 - Acceptable	4 - More than Acceptable	5 - Outstanding
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Candidate 1 - Situation 1**

Benchmarks

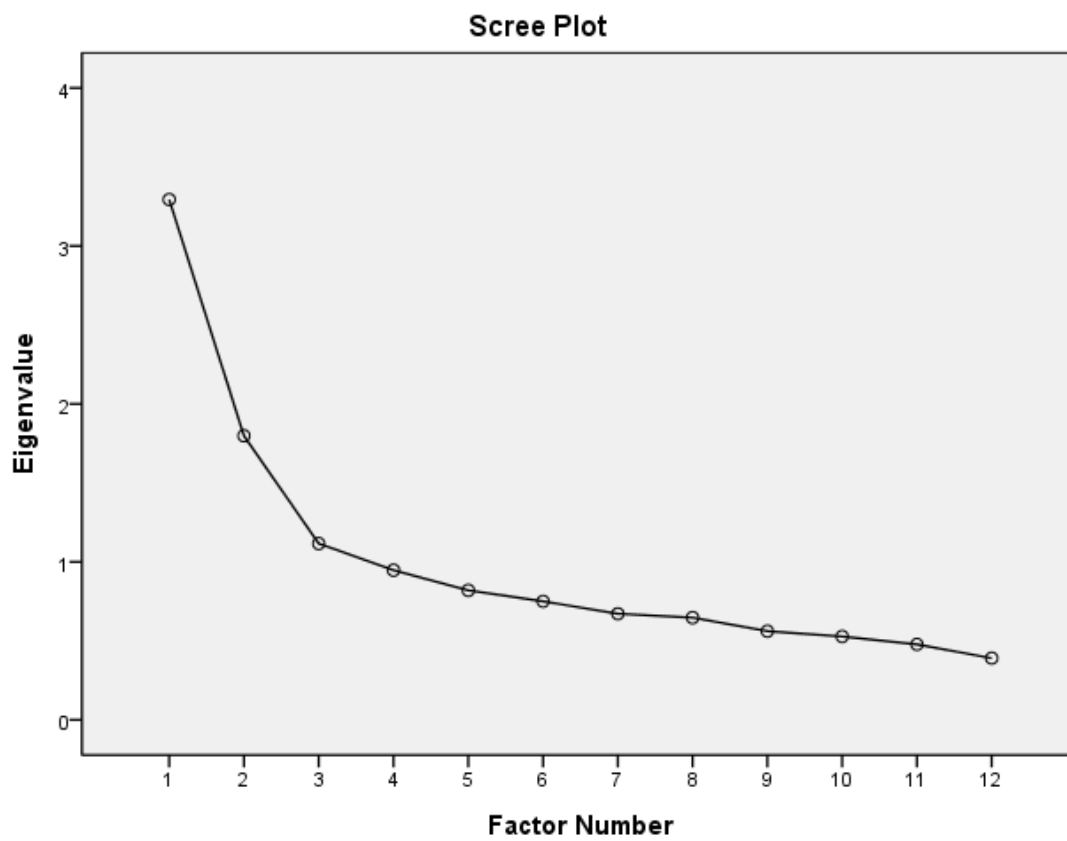
**Outstanding = 5**  
Candidate identifies several alternative methods for effectively dealing with the situation. For example, he/she would discuss the issue directly with the co-worker AND would seek advice from peers/supervisor

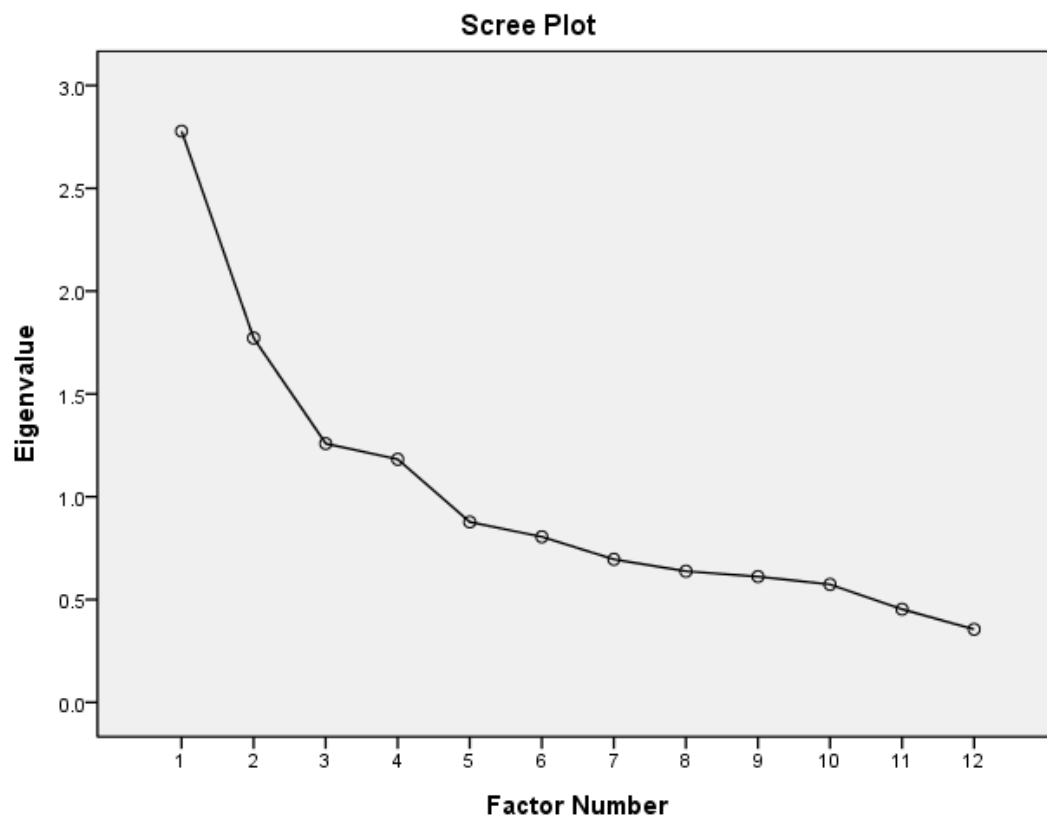
**More Than Acceptable = 4**  
Candidate states he/she would speak with the co-worker in private in a tactful manner about how the behavior may be bothering others who are helping cook and clean in the kitchen  
Candidate emphasizes the importance of talking to the co-worker in a tactful manner about the issue

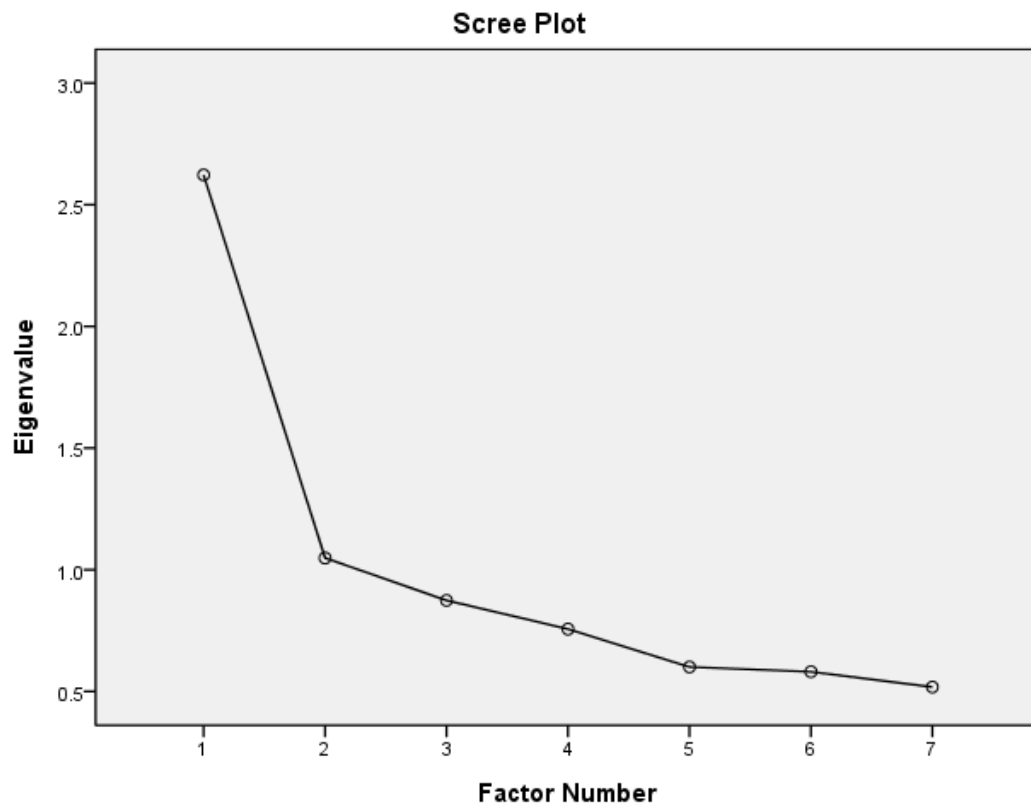
**Acceptable = 3**  
Candidate states he/she would ask for a supervisor's assistance with the co-worker before talking directly to the co-worker  
Candidate states he/she would privately explain to the co-worker that he is not doing his fair share of cleaning  
Candidate states he/she would ask a more senior firefighter to handle the situation

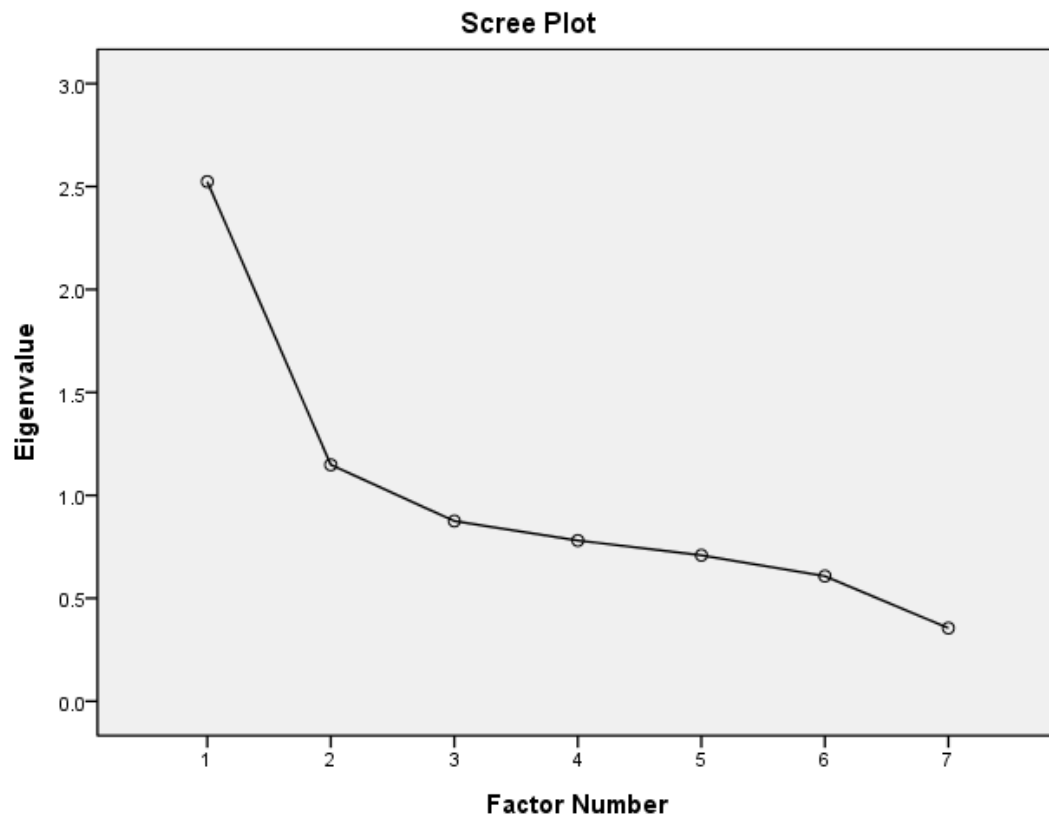
**Less Than Acceptable = 2**  
Candidate states that he/she may not be able to change the co-worker  
Candidate states he/she would not talk to anyone else about it because it is a private matter  
Candidate states he/she would confront the co-worker about the situation in front of other firefighters  
Candidate states he/she would leave a note for the co-worker asking him start doing his part of the cleaning after meals

**Unacceptable = 1**  
Candidate states he/she would make fun of the co-worker in front of other firefighters  
Candidate states he/she would retaliate in secret against the co-worker  
Candidate states he/she would threaten the co-worker with retaliation if he does not start doing his part of the cleaning

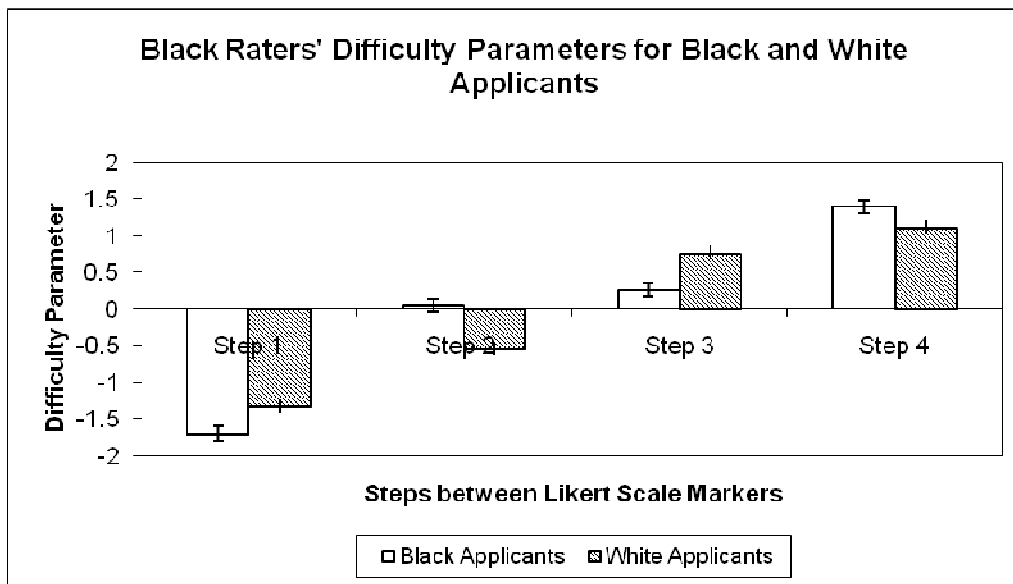
**Figure 4.1 Scree Plot for Black-Referent Items, Student Sample (Reverse Included)**

**Figure 4.2** Scree Plot for White-Referent Items, Student Sample (Reverse Included)

**Figure 5.1 Scree Plot for Black-Referent Items, Student Sample (Reverse Excluded)**

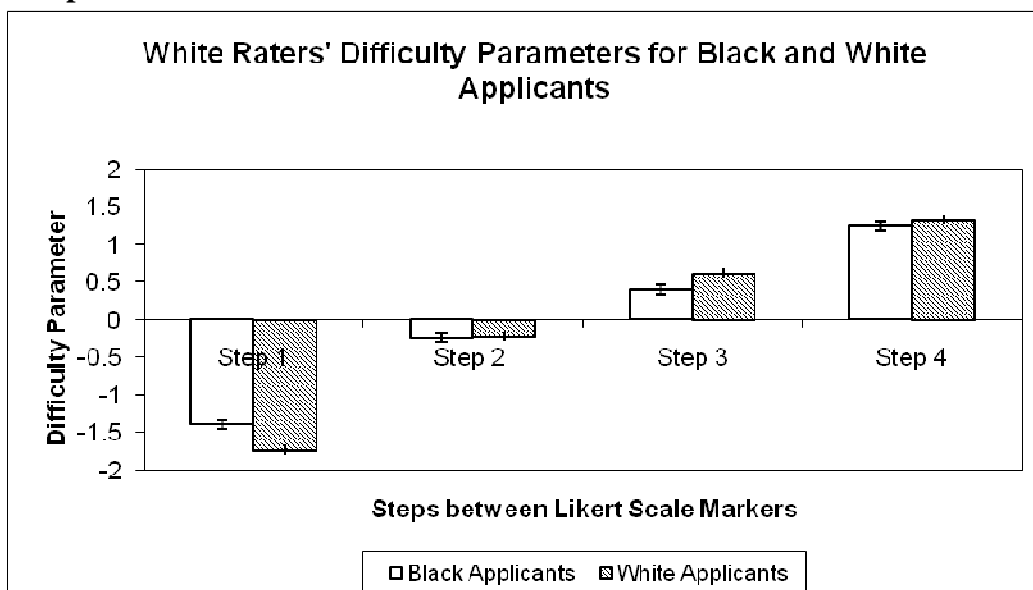
**Figure 5.2 Scree Plot for White-Referent Items, Student Sample (Reverse Excluded)**

**Figure 6.1 Black Raters' Assessment of Black and White Applicants – Student Sample**



*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*

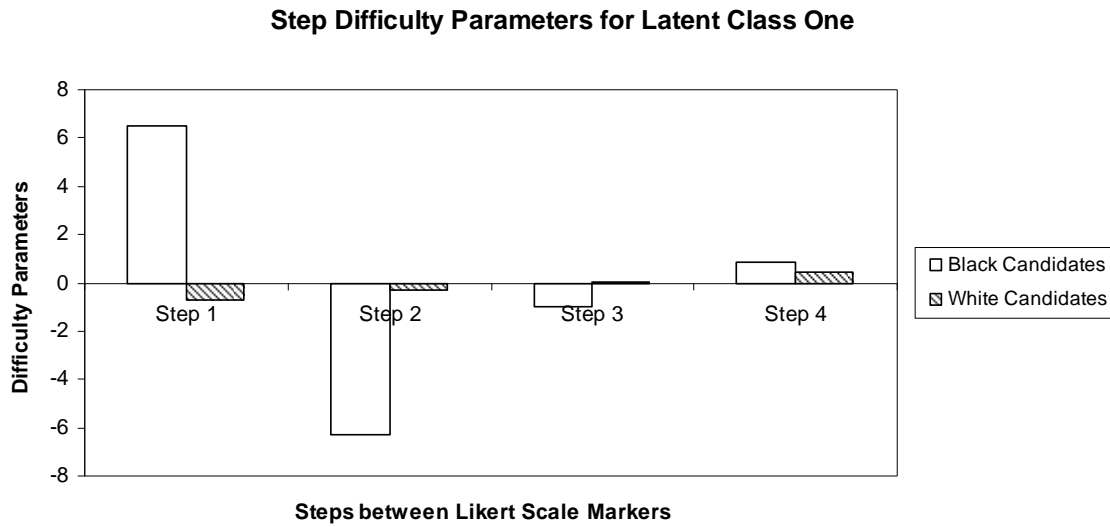
**Figure 6.2 White Raters' Assessment of Black and White Applicants – Student Sample**



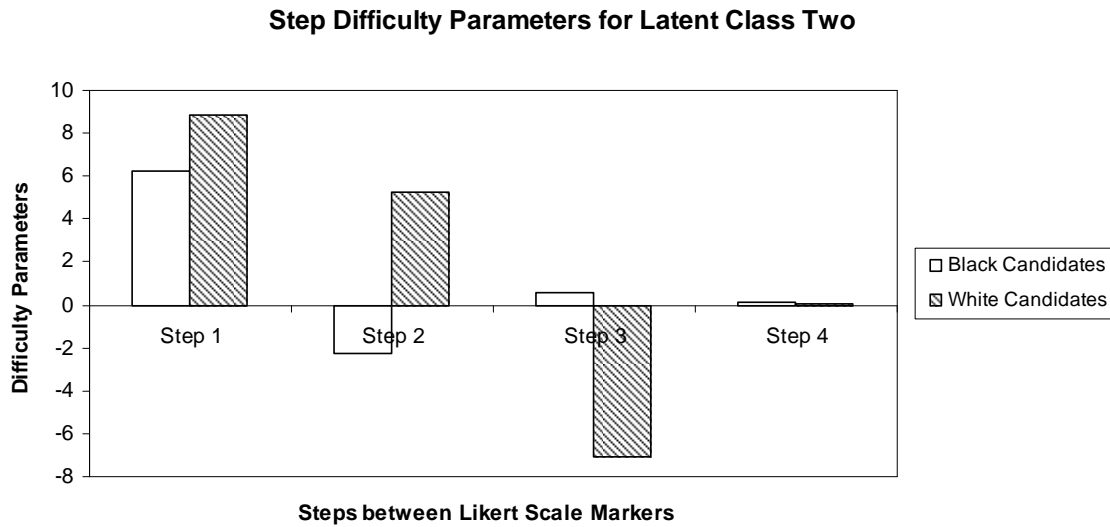
*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*



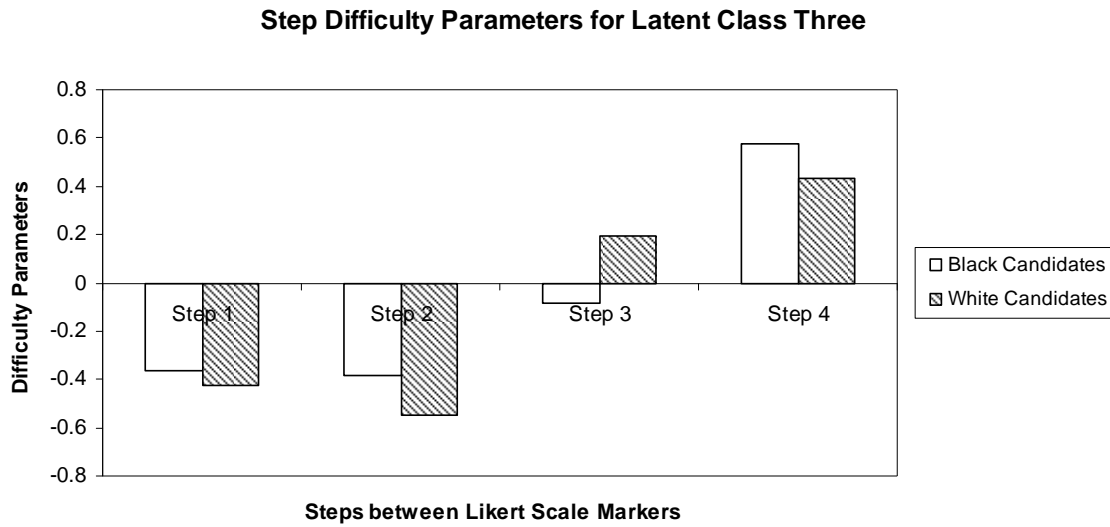
**Figure 7.1 Ratings of Black and White Candidates – Latent Class One**



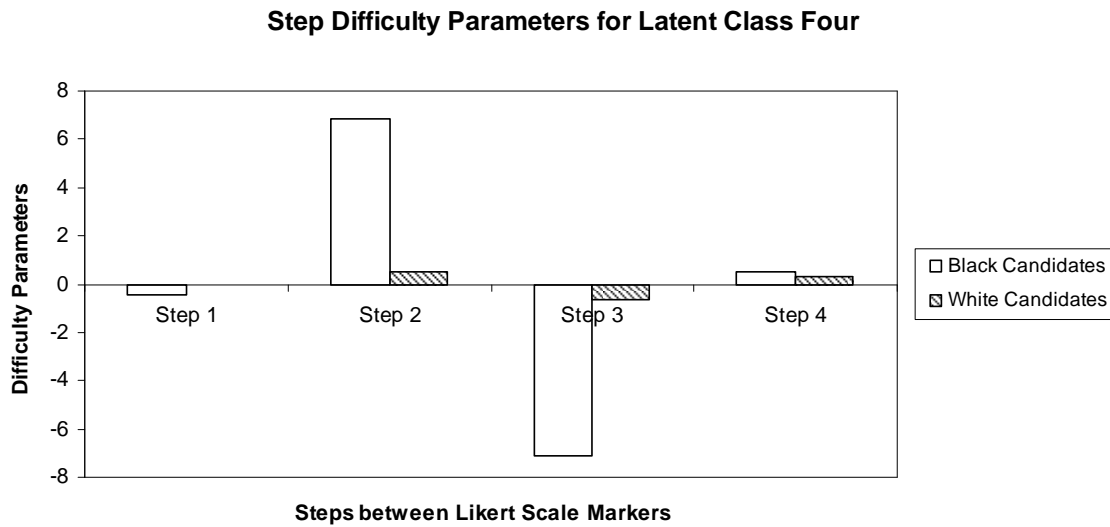
**Figure 7.2 Ratings of Black and White Candidates – Latent Class Two**

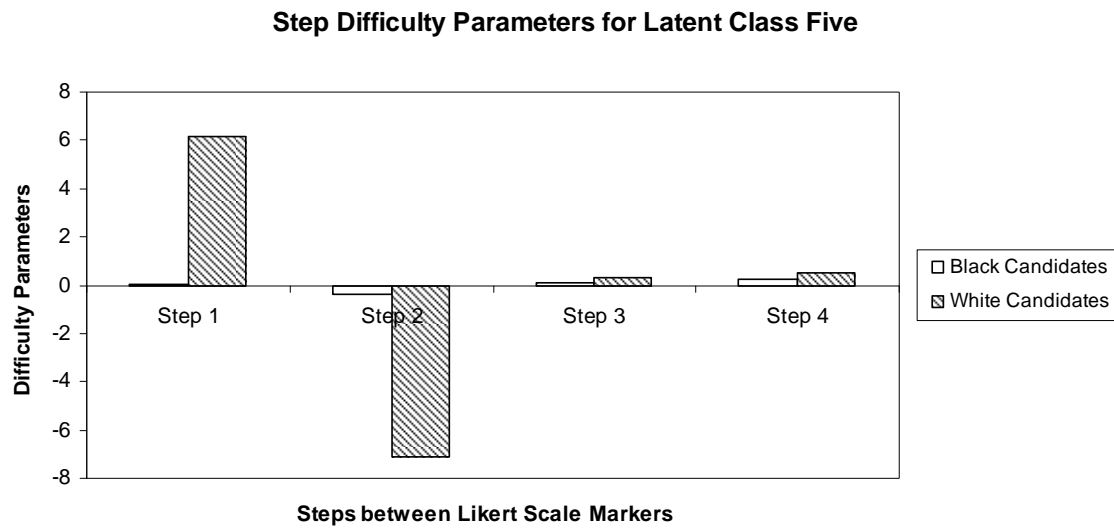


**Figure 7.3 Ratings of Black and White Candidates – Latent Class Three**

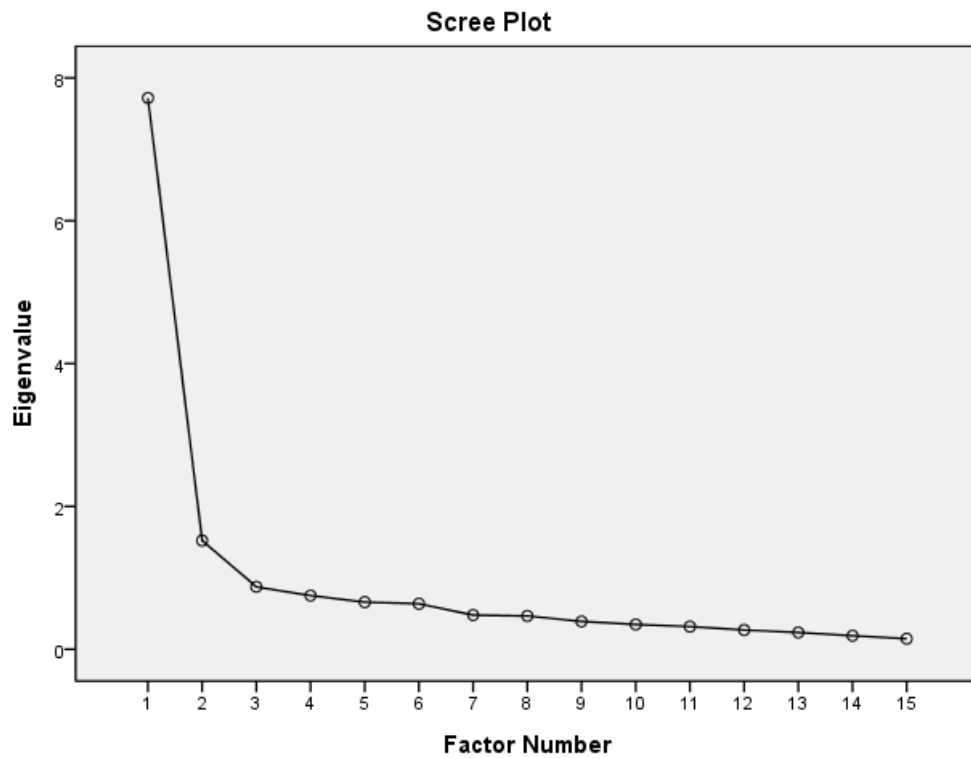


**Figure 7.4 Ratings of Black and White Candidates – Latent Class Four**

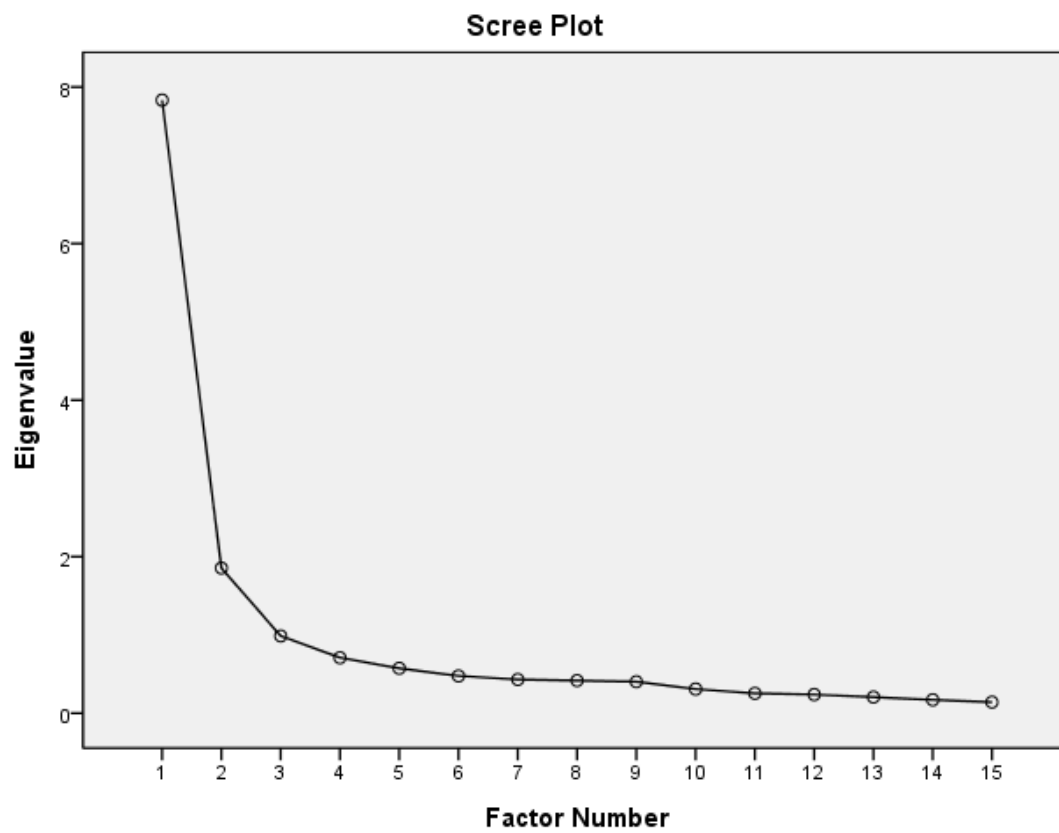


**Figure 7.5 Ratings of Black and White Candidates – Latent Class Five**

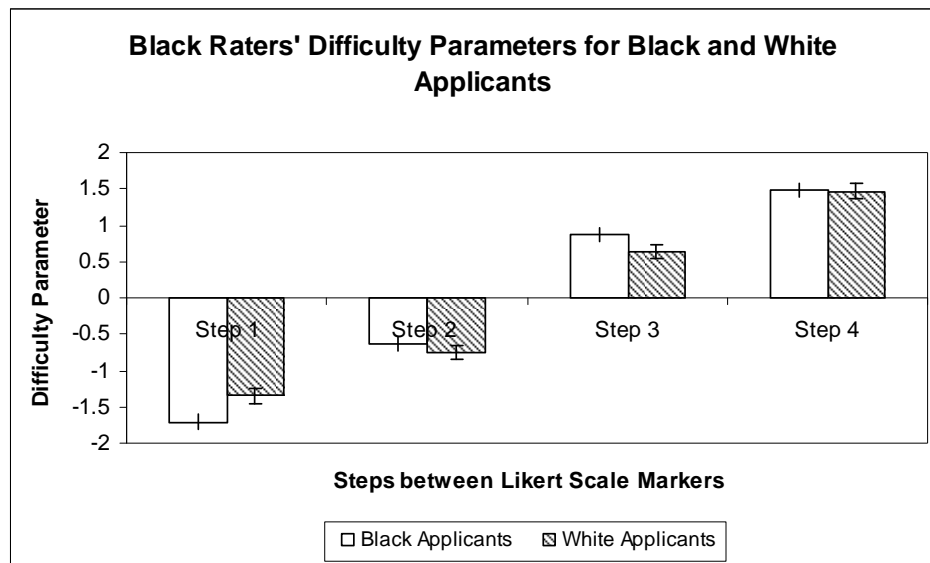
**Figure 8.1** Scree Plot for Black-Referent Items, Adult Sample (Second Prejudice Scale)



**Figure 8.2** Scree Plot for White-Referent Items, Adult Sample (Second Prejudice Scale)

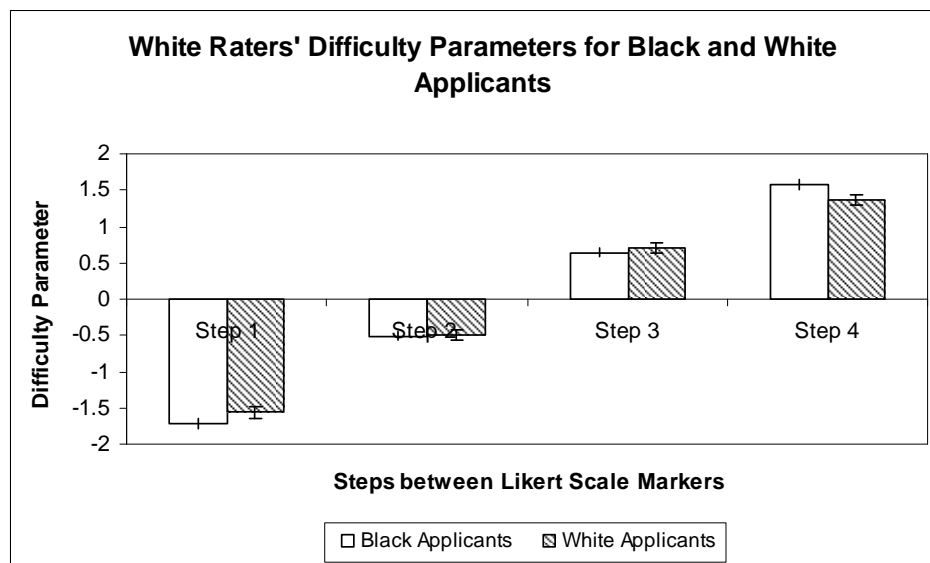


**Figure 9.1 Black Raters' Assessment of Black and White Applicants – Adult Sample**



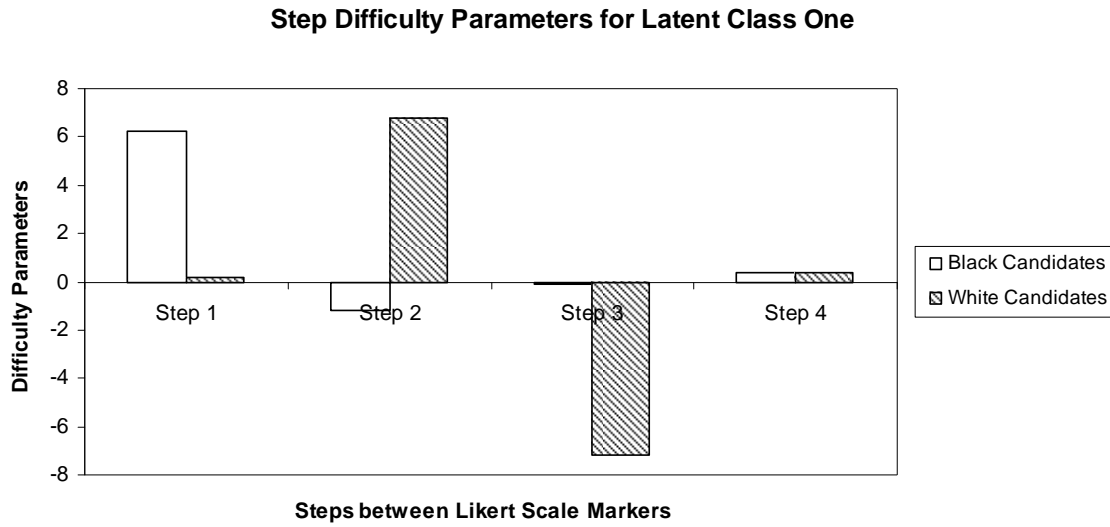
*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*

**Figure 9.2 White Raters' Assessment of Black and White Applicants – Adult Sample**

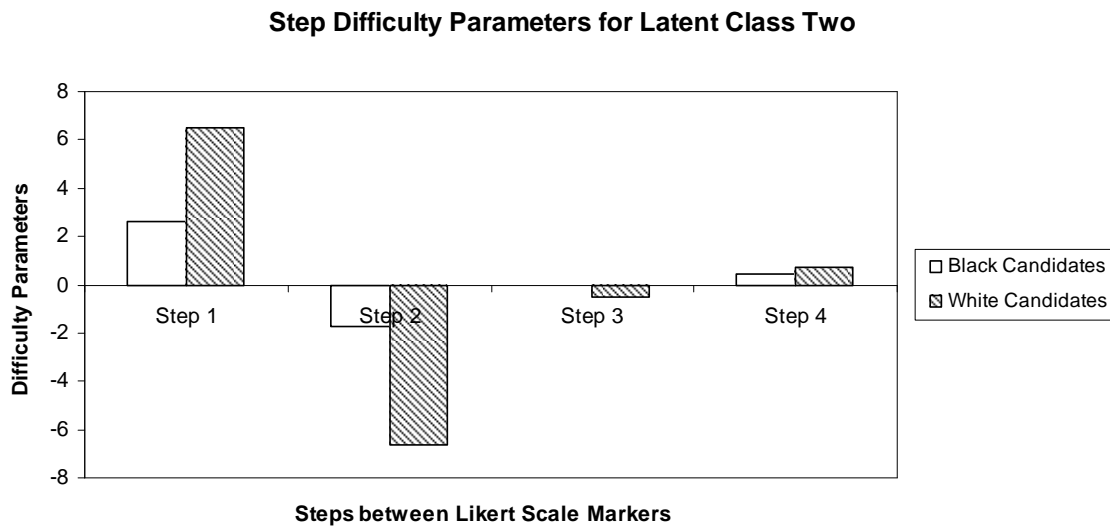


*Note: There are no standard errors for Step 4, since these parameters are fixed. For graphical purposes only, error bars were generated using average standard errors between Steps 1-3.*

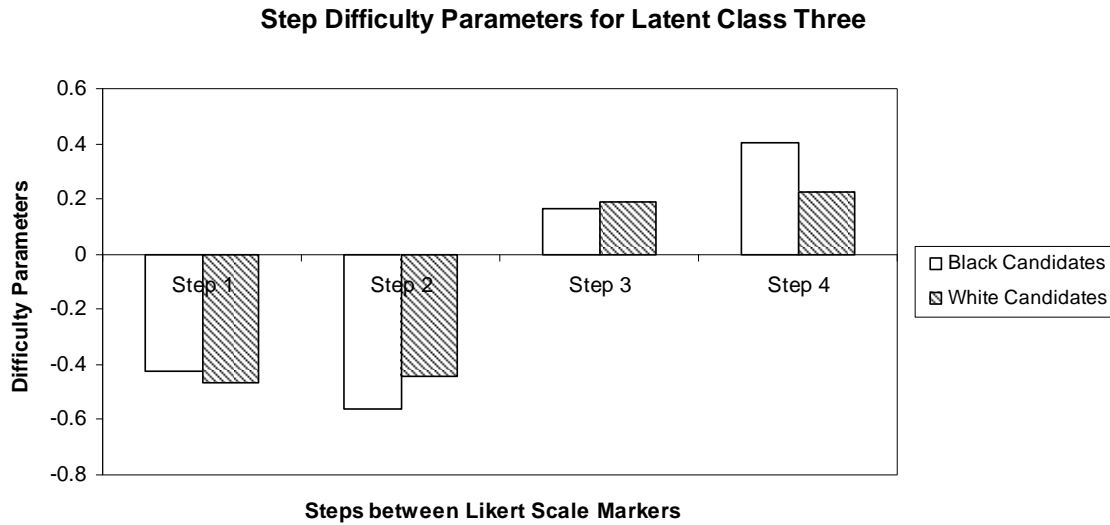
**Figure 10.1 Ratings of Black and White Candidates – Latent Class One, Adult Sample**



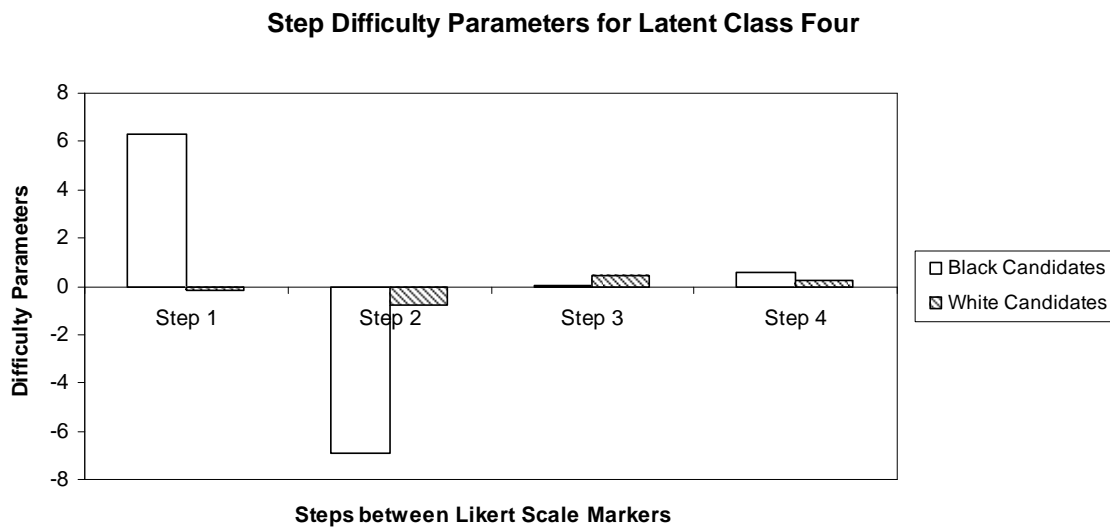
**Figure 10.2 Ratings of Black and White Candidates – Latent Class Two, Adult Sample**



**Figure 10.3 Ratings of Black and White Candidates – Latent Class Three, Adult Sample**

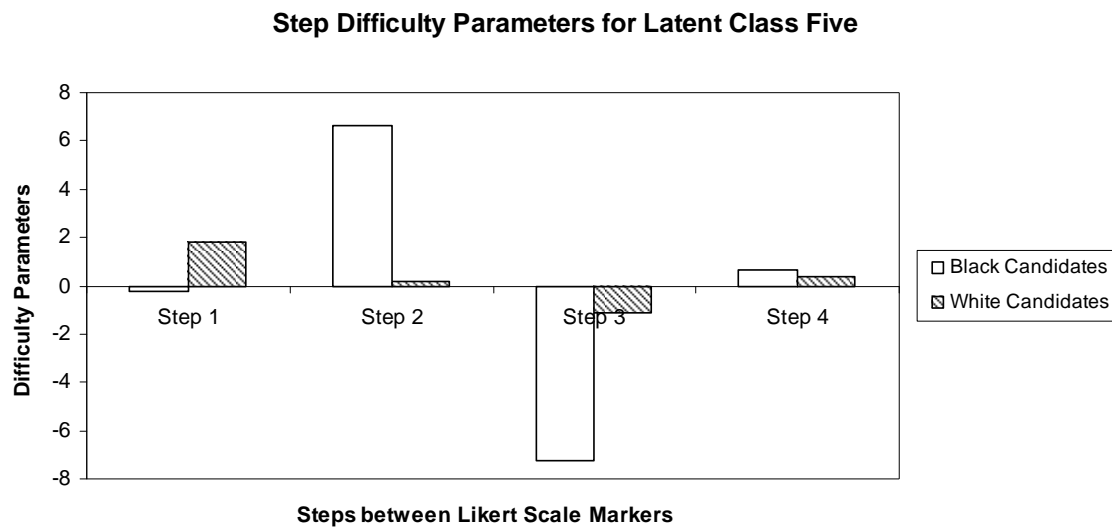


**Figure 10.4 Ratings of Black and White Candidates – Latent Class Four, Adult Sample**





**Figure 10.5 Ratings of Black and White Candidates – Latent Class Five, Adult Sample**



## References

- Abwender, D. A. & Hough, K. (2001). Interactive effects of characteristics of defendant and mock juror on U.S. participants' judgment and sentencing recommendations. *The Journal of Social Psychology, 141*, 603-615.
- Alvesson, M. & Billing, Y. D. (1992). Gender and organization: Towards a differentiated understanding. *Organization Studies, 13*, 73-102.
- Baltes, B. B., Bauer, C. B., & Frensch, P. A. (2007). Does a structured free recall intervention reduce the effect of stereotypes on performance ratings and by what cognitive mechanism? *Journal of Applied Psychology, 92*, 151-164.
- Bass, A. R. & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology, 57*, 101-109.
- Bauer, C. C. & Baltes, B. B. (2002). Reducing the effects of gender stereotypes on performance evaluations. *Sex Roles, 47*, 465-476.
- Bernardin, H. J. & Buckley, M. R. (1981). Strategies in rater training. *The Academy of Management Review, 6*, 205-212.
- Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307-324.
- Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues, 55*, 429-444.
- Brewer, M. B. (2001). Ingroup identification and intergroup conflict: When does ingroup love become outgroup hate? In R. D. Ashmore, Jussim, L., & Wilder, D. (Eds.), *Social identity, intergroup conflict, and conflict reduction* (Vol. 3, pp. 17-41). New York: Oxford University Press.

- Brewer, M. B. (2007). The importance of being *we*: Human nature and intergroup relations. *American Psychologist*, *62*, 728-738.
- Brigham, J. C. (1971). Ethnic Stereotypes. *Psychological Bulletin*, *76*, 15-38.
- Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology*, *23*, 1933-1967.
- Brown, R. (2000) Social identity theory: past achievements, current problems, and future challenges. *European Journal of Social Psychology*, *30*, 745-778.
- Butt, D. S. & Signor, E. I. (1976). Social images of disadvantaged groups. *Social Behavior and Personality*, *4*, 145-151.
- Chaiken, S. & Eagly, A. H. (1983). Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology*, *45*, 241-256.
- Chatman, C. M. & von Hippel, W. (2001). Attributional mediation of in-group bias. *Journal of Experimental Social Psychology*, *37*, 267-272.
- Chattopadhyay, P. (1999). Beyond direct and symmetrical effects: The influence of demographic dissimilarity on organizational citizenship behavior. *The Academy of Management Journal*, *42*, 273-287.
- Clark, K. B. & Clark, M. K. (1947). Racial identification and preference in Negro children. In T. Newcomb & Hartley, E. (Eds.), *Readings in Social Psychology* (pp 169-178). New York: Holt.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5-18.

- Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*, 315-319.
- DuBios, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology, 78*, 205-211.
- Echebarria-Echabe, A. & Guede, E. F. (2007). A new measure of anti-Arab prejudice: reliability and validity evidence. *Journal of Applied Social Psychology, 37*, 1077-1091.
- Ellemers, N., Wilke, H., & Van Knippenberg, A. (1993). Effects of the legitimacy of low group or individual status on individual and collective status-enhancement strategies. *Journal of Personality and Social Psychology, 64*, 766-778.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- 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.
- Foti, R. J., Knee, R. E., & Backert, R. S. (2008). Multi-level implications of framing leadership perceptions as a dynamic process. *The Leadership Quarterly, 19*, 178-194.
- Ford, J. K., Kraiger, K., & Schechtman, S. L. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin, 99*, 330-337.

- Frazer, R. A., & Wiersma, U. J. (2001). Prejudice vs. discrimination in the employment interview: We may hire equally, but our memories harbour prejudice. *Human Relations, 54*, 173-191.
- 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.
- Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal, 33*, 64-86.
- Grove, D. A. (1981). A behavioral consistency approach to decision making in employment selection. *Personnel Psychology, 34*, 55-64.
- Hamner, W. C., Kim, J. S., Baird, L., & Bigoness W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59*, 705-711.
- Hanges, P. J., Braverman, E. P., & Rentsch, J. R. (1991). Changes in raters' perceptions of subordinates: A catastrophe model. *Journal of Applied Psychology, 76*, 878-888.
- Hau, K. T. & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology, 57*, 327-351
- Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry, 19*, 49-71.

- Henry, P. J. & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23*, 253-283.
- Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup Bias. *Annual Review of Psychology, 53*, 575-604.
- Hogg, M. A., & Abrams, D. (1988). *Social identifications: A social psychology of intergroup relations and group processes*. New York: Routledge.
- Huffcut, A. I. & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*, 179-189.
- Johnson, J. D. & Lecci, L. (2003). Assessing anti-white attitudes and predicting perceived racism: The Johnson-Lecci scale. *Personality and Social Psychology Bulletin, 29*, 299-312.
- Jost, J. T. & Burgess, D. (2000). Attitudinal ambivalence and conflict between group and system justification motives in low status groups. *Personality and Social Psychology Bulletin, 26*, 293-305.
- Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes : A comparison and integration of three theories. *Journal of Personality and Social Psychology, 52*, 536-546.
- King, E. B., Madera, J. M., Hebl, M. R., Knight, J. L., & Mendoza, S. A. (2006). What's in a name? A multiracial investigation of the role of occupational stereotypes in selection decisions. *Journal of Applied Social Psychology, 36*, 1145-1159.
- Lepore, L. & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275-287.

- Lin, M. H., Kwan, V. S. Y., Cheung, A., & Fiske, S. T. (2005). Stereotype content model explains prejudice for an envied outgroup: Scale of anti-Asian American stereotypes. *Personality and Social Psychology Bulletin, 31*, 34-47.
- Lipponen, J. & Leskinen, J. (2006). Conditions of Contact, Common In-Group Identity, and In-Group Bias toward Contingent Workers. *The Journal of Social Psychology, 146*, 671-684.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
- Masters, G. N. & Wright, B. D. (1996). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer
- McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In Dovidio, J. F. & S. L. Gaertner (Eds), *Prejudice, discrimination, and racism* (pp. 91-125), Orlando, FL: Academic Press.
- McKay, P. F. & McDaniel, M. A. (2006). A reexamination of black-white mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91*, 538-554.
- Meulders, M. & Xie, Y. (2004). Person-by-Item Predictors. In P. De Boeck & Wilson, M. (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 213-240). New York, NY: Springer-Verlag LLC.
- Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field study of adverse impact and generalizability. *Academy of Management Journal, 25*, 598-606.

- Mullins, T. W. (1982). Interviewer decisions as a function of applicant race, applicant quality, and interviewer prejudice. *Personnel Psychology*, *35*, 163-174.
- Nowak, A., Vallacher, R. R., Tesser, A., & Borkowski, W. (2000). Society of self: The emergence of collective properties in self-structure. *Psychological Review*, *107*, 39-61.
- Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3<sup>rd</sup> edition). London: McGraw Hill.
- Otten, S. & Wentura, D. (1999). About the impact of automaticity in the Minimal Group Paradigm: Evidence from affective priming tasks. *European Journal of Social Psychology*, *29*, 1049-1071.
- Parsons, C. K. & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, *69*, 557-568.
- Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology*, *59*, 475-486.
- Perrin, B. M., Barnett, B. J., Walrath, L., Grossman, J. D. (2001). Information order and outcome framing: An assessment of judgment bias in a naturalistic decision-making context. *Human Factors: The Journal of the Human Factors*, *43*, 227-238.
- Peterson, L. & Blank, H. (2003). Ingroup bias in the minimal group paradigm shown by three-person groups with high or low state self-esteem. *European Journal of Social Psychology*, *33*, 149-162.



- Petty, R. E., Wegener, D. T. (1993). Flexible correction processes in social judgment: correcting for context induced contrast. *Journal of Experimental Social Psychology, 29*, 137–165.
- Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research, 7*, 156–176.
- Plant, E. A. & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811-832.
- Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38*, 76-91.
- Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74*, 770-780.
- Ross, J., Irani, I., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. *CHI EA*, 2863-2872.
- Roth, P., Bobko, P., McFarland, L., & 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.
- Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

- Scheepers, D., Spears, R., Doosje, B., & Manstead, A. S. R. (2006). Diversity in in-group bias: Structural factors, situational features, and social functions. *Journal of Personality and Social Psychology, 90*, 944-960.
- Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746.
- Schneider, D. J. (2004). *The psychology of stereotyping*. New York, NY: The Guilford Press.
- Schroyens, W., Shaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1713-1734.
- Schutz, H. & Six, B. (1996). How strong is the relationship between prejudice and discrimination? A meta-analytic answer. *International Journal of Intercultural Relations, 20*, 441-462.
- Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management, 23*, 659-678.
- Stangor, C., Sullivan, L. A., & Ford, T. E. (1991). Affective and cognitive determinants of prejudice. *Social Cognition, 9*, 359-380.
- Stangor, C. (2009). The study of stereotyping, prejudice, and discrimination within social psychology: A quick history of theory and research. In Nelson, T. D. (Ed.) (2009). *Handbook of prejudice, stereotyping, and discrimination* (pp. 1-22). New York, NY: Taylor and Francis Group.

- Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Brooks/Cole.
- Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimizing racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social Justice Research, 21*, 263-296.
- Tsui, A. S., Egan, T. D., & O'Reilly, C. A. (1992). Being Different: Relational Demography and Organizational Attachment. *Administrative Science Quarterly, 37*, 549-579.
- Turner, J., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A social categorization theory*. Oxford, England: Blackwell.
- Vanbeselaere, N. (1993). Ingroup bias in the minimal group situation: An experimental test of the inequity prevention hypothesis. *Basic and Applied Social Psychology, 14*, 385-400.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.
- Waldman, D. A. & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 897-901*.
- Wang, M. & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods, 14*, 24-31.

- Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, *68*, 36-51.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.
- Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology*, *67*, 189-205.
- Ziegert, J. C. & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, *90*, 553-562.
- Ziegler, R., Arnold, F., & Diehl, M. (2007). Communication modality and biased processing: A study on the occasion of the German 2002 election TV debate. *Basic and Applied Social Psychology*, *29*, 175-184.

---

<sup>i</sup> Results from Study 3 regarding the differences between the scripts obtained from black- and white-candidates were also significant ( $t(183) = -2.74, p < 0.05$ ), with the script taken from the black candidate receiving a higher rating than the script taken from the white candidate.

<sup>ii</sup> Results from Study 3 regarding the impact of design factors on ratings replicated the Study 2 findings. That is, candidate vocal profiles, attractiveness, masculinity, and candidate order impacted ratings. For a summary of these findings, refer to Tables 41 and 42.