When Group Identities Matter
Bias in Performance Appraisal

LORIANN ROBERSON
Teachers College, Columbia University

BENJAMIN M. GALVIN
Arizona State University

ATIRA CHERISE CHARLES
Arizona State University

Abstract
Performance appraisals are a critical part of organizational life, and bias in appraisals is consistently mentioned as a barrier to advancement for diverse workers. This chapter reviews the literature on rater bias in performance appraisals, defined as effects on performance ratings due to ratee category membership. We focus on the major theoretical frameworks (e.g., stereotype fit and relational demography) used in the study of bias, organizing the research findings by the four most commonly studied demographic categories: gender, race, age, and disability. The review allows better understanding of the gaps in our knowledge and identifies needed future research directions in this literature stream. We conclude the chapter with several concerns including a lack of empirical research testing propositions concerning the effect of organizational variables on bias.

Introduction
Organizational diversity and equal employment opportunity (EEO) initiatives have emphasized entry and recruitment as a first step in creating a diverse work force (Avery & McKay, 2006). The success of these efforts can perhaps most
dramatically be seen in the changes in rates of entry of women and people of color into managerial and professional positions. For example, the number of African American women entering management positions increased by 79% from 1986 to 1996 (Combs, 2003). Recent figures show that the representation of women and people of color in managerial and professional positions is becoming closer to their representation in the labor force (Catalyst, Inc., 2006).

But still, disparities remain, particularly in rates of advancement into top leadership positions. A recent study by Catalyst, Inc. (2006) found that 37% of managers, but only 16% of corporate officers, were female and that the rate of growth for women into top positions has slowed in the last three years. A similar picture was found for people, particularly women, of color. Other studies have documented slower rates of advancement for men and women of color than for their White peers, even for those who eventually reach upper echelons (Bell & Nkomo, 2001; D. A. Thomas & Gabarro, 1999). Thus, a ceiling—be it glass, concrete (Catalyst, Inc., 1999), or bamboo (Hyun, 2006)—continues to be a reality of organizational life. Attention has shifted from organizational entry to advancement as the most important diversity challenge facing organizations (Kilian, Hukai, & McCarty, 2005). Disparate advancement rates mean limited career paths for women and minorities and are believed to contribute to higher turnover for members of these groups (Kilian et al., 2005).

A number of possible reasons for these disparities have been identified including a lack of mentors or early job challenge, and exclusion from informal networks (Kilian et al., 2005). But also included in the list of probable causes is a set of factors related to performance evaluation and appraisal such as stereotypes, low expectations, and double standards (Catalyst, Inc., 2006; Kilian et al., 2005). For example, 41% of African American female managers surveyed by Bell and Nkomo (2001) expressed concern with double standards, saying that they had to perform better than their male colleagues had to. Concern with appraisal bias as a potential contributor to outcome disparities (Dipboye, 1985; Nieva & Gutek, 1980) has spurred and maintained interest in research on this issue and interest in the development of theories to explain when, why, and how bias in appraisal can occur.

What is bias in appraisal? Finding greater disparities between the representation of women and men or of Whites and people of color at upper management levels does not demonstrate unequivocally that performance evaluations (on which promotions are often based) are biased. Nor do employee concerns about discrimination in appraisals and promotions; nor does the existence of mean differences (e.g., women vs. men; Black vs. White) in the appraisals themselves. While ideally we might hope to see no differences in outcomes by group memberships, observed differences in level of attainment or performance could be the result of differences in skills, abilities, education, or experience. Fiske (2002) defined bias as “reacting to a person on the basis of perceived membership in a single human category, ignoring other category
attributes" (p. 123). In terms of performance appraisal, this means a systematic effect due to category membership (e.g., gender, race, age, disability), unrelated to actual performance (Colella, DeNisi, & Varma, 1998; Ilgen & Youtz, 1986; Nieva & Gutek, 1980; Stauffer & Buckley, 2005). Finding this effect in research can be accomplished (or attempted) in several ways. In the laboratory, the actual work behaviors and outputs seen by evaluators can be equated. In the field, other major contributors to performance (e.g., ability or experience) or objective performance level can be statistically controlled. Some studies use within-subjects designs, so that ratees and performance are constant across raters. When performance or its other causes are controlled, and effects due to category membership remain—when group identities still matter—bias is said to exist.

This chapter reviews literature on bias in performance appraisals. In reviewing the literature, we concentrated on management and applied psychology journals, but we included work in basic or social psychology that considered performance evaluations in organizational contexts. Our review is structured around the major theoretical frameworks used to study and explain bias in the management literature. We consider work for four commonly studied demographic categories: gender, race, age, and disability. While we focus on the theories of bias for each of these categories, we also briefly summarize recent (i.e., post-1990) results of empirical studies based on these perspectives. Finally, we reflect on the state of our knowledge and the field, and offer suggestions for future research.

History and Overview

Because of their importance for administrative decisions and employee development, as well as their perceived vulnerability to error and inaccuracy (Landy & Farr, 1980), performance evaluations have been a subject of study since the early days of industrial and organizational psychology. Research was conducted on rating errors, formats, and rater and ratee demographics (mostly race, sex, and age) as sources of rating variance. Thus, the study and reporting of group differences in appraisals, especially for gender and race, have a long history. The 1980 review of Landy and Farr reported two common findings from research. First, in occupations perceived as masculine, ratings of females tended to be lower than males. This pattern suggested that gender bias may be a function of role incongruity (Nieva & Gutek, 1980). Second, raters were often found to give more favorable evaluations to same-race ratees. This pattern received less explanation. As noted by Ilgen and Youtz (1986), in the study of race and performance appraisal, scholars seemed content to document the existence of differences rather than seek to explain why they might be occurring. The rise of the cognitive models of person perception in social psychology and their application to the performance appraisal process set the stage for more theoretical approaches to the study of bias in appraisal.
Using a cognitive approach, Heilman (1983) presented the lack-of-fit model to explain the pattern of findings for gender. Tsui and O’Reilly (1989) introduced the relational demography approach, providing a theoretical basis for understanding similarity effects on ratings. These two papers formed the foundation for the two major theoretical approaches to the study of bias today. In the last 15 to 20 years, work within these frameworks on gender and race bias has continued, and models of age and disability bias have appeared as well.

Models of bias utilize either or both of the two theoretical frameworks. One (perhaps the predominant) theoretical model is what we will refer to as “stereotype fit” (Dipboye, 1985). This type of model proposes a matching process such that evaluators compare their stereotype (e.g., the beliefs about the attributes or characteristics associated with an identity group; Dovidio & Hebl, 2005) of a ratee with another stereotype of the perceived requirements of the ratee’s job or role. These models argue that some jobs and roles become “typed,” in that the attributes believed necessary for success are those associated with a particular group. Bias is predicted to occur when there is little fit between the two, and the stereotype of a ratee’s group is incongruous with the stereotype of his or her job. Incongruity is hypothesized to affect performance evaluations negatively through expectations of success. Theories based largely on stereotype fit models have been proposed to explain both selection and appraisal bias for gender (e.g., Eagly & Karau, 2002; Heilman, 1983, 2001; Perry, 1997; Perry, Blake-Davis, & Kulik, 1994), age (e.g., Perry, 1997; Perry & Finkelstein, 1999) and disability (e.g., Colella, DeNisi, & Varma, 1997; D. L. Stone & Colella, 1996).

The second major theoretical framework used to predict bias in appraisal comes from relational demography. The relational demography approach investigates the effects of an individual’s demographic similarity-dissimilarity relative to others in a dyad or group (Tsui & Gutek, 1999). This model proposes that demographic dissimilarity of a ratee from his or her rater or work group will result in less favorable performance evaluations. Researchers have used the demography framework to explain appraisal bias for gender (Tsui & O’Reilly, 1989), race (Mount, Sytsma, Hazucha, & Holt, 1997), age (Tsui, Porter, & Egan, 2002), and disability (Colella et al., 1997).

These two theoretical frameworks (stereotype fit and relational demography) specify the general conditions under which bias will occur (an incongruous job or a dissimilar rater/work group) and, thus, specify where researchers should investigate bias. Within these general frameworks, researchers have studied a number of other theories of smaller scope that seek to explain in greater detail the process through which bias is manifested in evaluations—how expectancies or dissimilarity influence ratings. These process models, developed in social psychology, are often invoked to explain or predict particular patterns of ratings.
We begin by discussing the general frameworks and briefly summarize the research evidence for the four demographic categories. Then, we discuss the process models and the role of individual and of contextual factors in bias.

The Stereotype Fit Framework
Models of bias relying on the stereotype fit framework predict that bias occurs when the stereotype of a ratee's group is incongruous with the stereotype of his or her job. As the lack of fit model for gender states, “[T]he presumed lack of fit that arises from a perceived attributes-job requirements incongruity underlies each of the many varieties of sex-bias encountered in the work world” (Heilman, 1983, p. 280). This framework rests on the assumption that both jobs or roles, and identity groups are stereotyped. Accordingly, research has documented the existence and content of stereotypes for gender, disability, and age groups (Colella, 1996; Heilman, 2001; Kulik, Perry, & Bourhis, 2000; Wittenbrink, Judd, & Park, 1997), as well as evidence for the sex and age typing of jobs (Perry & Bourhis, 1998; Saks & Waldman, 1998). Research on disability has shown that people hold different expectations of success for various disability job categories (Colella, 1996).

Stereotype fit models also rely on a large body of social cognition research to predict that incongruity between the job and ratee group stereotypes negatively affects performance evaluations through expectations of success. Stereotypes create expectations about the likely characteristics and behaviors of group members (Higgins & Bargh, 1987). Social cognition research has shown how stereotype-based expectancies can bias the perception, processing, and recall of information (Dovidio & Hebl, 2005). For example, in the absence of information perceivers fill in details consistent with their expectations (Kulik & Bainbridge, 2005). Research has also tested the central hypothesis of the fit model: that bias in evaluation for a particular demographic group depends on the typing of the job, typically indicated by a target demographic by job interaction (Kulik & Bainbridge, 2005). Although most of this research has focused on evaluations of job applicants in selection contexts, evidence supporting the hypothesis for performance evaluations has been found for several demographic categories. Much of the strongest evidence in support of the model concerns gender (Davison & Burke, 2000; Eagly & Karau, 2002), the category with which job congruity effects on appraisals were first noted. As predicted, women tend to be evaluated as less effective than men are in male-dominated leader roles and roles perceived as masculine. Women tend to be perceived as more effective than men are in roles perceived as feminine (Eagly, Karau, & Makhijani, 1995).

Models of age bias (Perry & Finkelstein, 1999) also incorporate the stereotype fit hypothesis, and much of the empirical research on age bias has been guided by this framework (Gordon & Arvey, 2004). Empirical results suggest that older ratees in “young-typed” jobs are penalized more than
younger ratees are in “old-typed” jobs. For example, Finkelstein, Burke, and Raju's (1995) meta-analysis of studies that tested the fit hypothesis revealed that younger workers received slightly higher ratings than older workers did for young-type jobs, but there was little difference in ratings of younger and older workers on old-type jobs. This pattern of results is consistent with some variations of the fit model specific to age. Lawrence (1988) proposed that age bias involved a matching process but argued that raters make comparisons not based on perceived attributes of the group and job, but based on the age of the ratee relative to perceived organizational norms for career progression (beliefs about the typical age of people in a role). Ratees whose age is similar to the norm are deemed “on schedule” and those younger than the norm are deemed “ahead of schedule.” Performance expectations and ratings will be positive for these individuals, particularly for those judged “ahead of schedule.” However, ratees older than the norm are deemed “behind schedule,” and this is predicted to have a negative impact on expectations and ratings. Maurer, Wrenn, and Weiss (2003) also proposed a fit process to predict appraisal bias affecting older and not younger workers. They suggested that the common stereotype that older workers are less motivated or able to learn (Chiu, Chan, Snape, & Redman, 2001) can negatively affect performance evaluations of older workers on jobs that are viewed as having a learning component.

D. L. Stone and Colella (1996) and Colella et al. (1997) also propose stereotype fit as one of the important processes affecting the evaluation of disabled ratees. The few empirical studies of the stereotype fit hypothesis for disability show mixed results. In Colella et al. (1998), negative bias for disabled ratees was seen under poor job fit conditions, but only on rankings and not on performance ratings. Colella and Varma (1999) also found that performance ratings were not affected by fit (or disability), but performance expectancies and reward recommendations were negatively affected for disabled ratees under poor fit conditions. In Lynch and Finkelstein (2006), fit had no impact on ratings of disabled ratees.

Given the popularity of the stereotype fit model, it is somewhat striking that it has not been used to explain race bias in appraisals. In their article examining factors affecting evaluations of minorities, Ilgen and Youtz (1986) argued that, although stereotype fit models appeared useful for explaining gender bias, they were less relevant when considering race bias, as jobs are not race typed to the extent that they are sex typed. However, status characteristics theory (Wagner & Berger, 1997) suggests that similar processes are involved in both race and gender bias. Gender and race are both viewed as diffuse status characteristics, which means that stereotypes about these groups contain beliefs about their status position and about member competence with regard to a wide range of abilities (Ridgeway, 2001). These assumptions about competence and social significance influence performance expectancies. Status characteristics theory proposes that these characteristics will be used to form...
expectancies when they are salient in the situation (i.e., in mixed groups, or when they are relevant to the task). Thus, the results of tests of the stereotype fit model for gender are also viewed as consistent with status characteristics theory (Ridgeway, 2001). If race is also associated with status, then race bias in appraisals should be seen on higher status roles or jobs. Although work in social psychology has utilized this perspective for explaining race bias, in the management literature most research on race bias in appraisals has been guided by the next framework.

The Relational Demography Framework

The relational demography approach investigates the effects of an individual’s demographic similarity/dissimilarity relative to others in a dyad, group, or organization (Riordan, Schaffer, & Stewart, 2005; Tsui & Gutek, 1999). This framework proposes that demographic dissimilarity of a ratee from a rater or work group will result in less favorable performance evaluations.

The relational demography framework relies on three theories to predict the effects of similarity and dissimilarity on processes and outcomes: social identity theory (Hogg & Abrams, 1988), self-categorization theory (Tajfel, 1982), and the similarity-attraction paradigm (Byrne, 1971). These theories explain the tendency to perceive others in terms of in-groups and out-groups and an individual’s attraction to and preference for similar others (Tsui, Egan, & O’Reilly, 1992). At the dyadic level, these theories predict that demographic dissimilarity of a ratee to a rater will result in less favorable performance evaluations, with liking or attraction as the mediator of these effects (Riordan, 2000). At the work group level, in addition to similarity-attraction, social identity, and social categorization theories, Kanter’s (1977) theory of tokenism is used to predict the effect of similarity/dissimilarity to the group on appraisals (Riordan et al., 2005). Kanter proposed that when members of one group make up 15% or less of the total, the salience of group memberships and associated stereotypes increases, and the behavior and performance of token individuals receive greater levels of scrutiny. The theoretical implication for appraisals is that individuals who are dissimilar to the group, particularly those who are members of low-status groups, are more likely to be given lower appraisals by the dominant group (Bettencourt & Bartholow, 1998). Here we will look at evidence for dyadic and group similarity effects on appraisals.

Dyadic Similarity

The more popular approach to investigating appraisal bias involves examining dyadic rater–ratee similarity and its effects on appraisals. Rater–ratee similarity effects have been reported for gender (Biernat, Crandall, Young, Kobrynowicz, & Halpin, 1998; Furnham & Stringfield, 2001; Tsui & O’Reilly, 1989; Tsui et al., 2002; Varma & Stroh, 2001). Results tend to show that gender similarity positively predicts performance ratings. There is some indication that
males and females may vary in the extent of in-group favoritism, although the direction of the relationship is unclear. In Biernat et al. (1998), male raters showed more in-group bias than female raters did, but in Tsui and O'Reilly, female raters showed greater similarity effects than males.

For age, results have been mixed. In Liden, Stilwell, and Ferris (1996) and Tsui and O'Reilly (1989), age similarity had no effect on performance ratings. Other findings suggest that the direction of the age difference may be important (Tsui & O'Reilly, 1989). Tsui et al. (2002) found that similarity in age between subordinates and employees was positively related to performance ratings and that older subordinates, relative to the supervisor, received the lowest performance scores. Shore, Cleveland, and Goldberg (2003) found predicted similarity effects on manager rated potential but not on performance ratings. On ratings, results opposite to those predicted were found. While older and younger managers rated the performance of younger employees similarly, older employees were rated more favorably by younger managers than they were by older managers.

In 1980, Landy and Farr noted race similarity effects on appraisals. In general, subsequent examinations of race have also reported rater-ratee similarity effects, with the bulk of studies examining Black and White groups (Biernat et al., 1998; Elvira & Town, 2001; Kraiger & Ford, 1985; Mount et al., 1997; Stauffer & Buckley, 2005). Tsui and O'Reilly (1989) found no evidence for race (Black-White) dissimilarity effects on performance ratings, but Tsui et al. (2002) reported that race dissimilarity affected ratings of extrarole performance. Several studies suggest a greater negative effect of racial dissimilarity on ratings made by White evaluators. In Mount et al. (1997), both Black and White superiors assigned more favorable ratings to same race raters, with the greatest difference between raters for Black ratees. Stauffer and Buckley (2005) found that higher ratings were given to White ratees by both White and Black supervisors. However, White supervisors rated Black ratees much lower than Black supervisors did. Biernat et al. (1998) reported that racial similarity effects were stronger for White raters than they were for non-White raters.

**Group Similarity**

Some studies of group similarity effects have utilized the compositional demography approach, which examines performance ratings as a function of the relative proportions of social identity groups in a work group or organizational unit. Studies of the effect of group composition on appraisals have been more common for gender than for other demographic variables (Reskin, McBrier, & Kmec, 1999), and the evidence is stronger. For example, Pazy and Oron (2001) explored the relationship between the proportion of women in military units and performance evaluation outcomes for high-ranking officers in the Israeli Defense Force. When women made up less than 10% of a unit, they received lower overall ratings than men did. In groups in which
women made up more than 10% of the unit, they generally received superior ratings to the men. In contrast, men's overall performance ratings did not vary with the proportion of women in the unit. Sackett, Du Bois, and Noe (1991) similarly reported that women received lower ratings than men did when they made up less than 20% of a group, but received higher evaluations than men did when they made up more than 50% of the group. Sackett et al. (1991) also examined the effect of race on work group composition on Black-White rating differences. The proportion of Blacks in the work group had no significant influence on ratings.

For age and disability, there have been few, if any, studies of the effects of group similarity on performance evaluations. Lawrence (1988) found that performance ratings were significantly related to deviations from age norms, not deviations from the age composition of those in the role. Lawrence concluded that, although demographic composition is often used as a proxy for age norms, the two are not identical and deviation from age norms is the important predictor of bias. However, age norms may be partly a function of age composition. A study of selection bias (Cleveland, Festa, & Montgomery, 1988) found that perceived age norms for the job and age typing of the job were related to the proportion of older individuals in the applicant pool. Similar to the findings of Lawrence, age composition was not related to evaluations when holding perceptions of age norms constant.

In summary, the stereotype fit and relational demography approaches provide broad frameworks that indicate where and when rating bias will occur, and thus where to look for it. Models of bias in management utilize either or both of these frameworks. For example, Colella et al.'s (1997) model of appraisal bias for disabled employees includes predictions derived from both stereotype fit and relational demography frameworks. The two frameworks are not inconsistent with each other. It has been proposed that the demographic composition of a job or role is one contribution to its typing (Perry, 1997), and the study of Cleveland et al. (1988) supports this for age. If so, the findings of a relationship of ratings for a particular group to its representation in the job are consistent with the stereotype fit and status characteristics predictions.

Evidence exists to support the central predictions of each framework, although overall, results are mixed. But mixed results are not surprising, as models of bias also propose a number of contextual and individual difference variables that can influence the extent to which bias occurs. Overwhelmingly, empirical research has not considered or measured proposed moderators, but rather tests basic hypotheses concerning congruency or similarity effects on evaluations. In addition, research has tended to define and to look for bias in appraisals in terms of lower, less favorable ratings for members of the target group. But as discussed in the next section, bias does not always mean lower ratings.
Process Theories

The stereotype fit and relational demography frameworks specify basic mechanisms through which bias occurs: through expectancies in the stereotype fit framework and through attraction/liking in the relational demography approach. However, within the theoretical bases of the two frameworks, other theories have been developed that provide greater detail on particular processes that are involved in bias. Here we review some of these process theories that explain in more detail how bias is manifested, and why "bias" does not always mean "lower ratings."

Performance Attributions

In both the stereotype fit and relational demography frameworks, performance attributions are one mechanism through which bias is hypothesized to occur. Within the stereotype fit framework, low expectations for success result in rater attribution of high performance to instable or external factors instead of internal factors such as ability. For example, models of gender bias argue that when a woman's level of performance is high, incongruity (which results in low expectations) results in denying women credit for their success (Eagly & Karau, 2002; Heilman, 2001). Heilman predicted this attribution pattern to be more likely in situations where individual contributions to success are ambiguous such as in team settings. Colella et al. (1997) also suggested that attributions contribute to appraisal bias for disabled ratees, arguing that raters would be expected to attribute poor performance to stable, uncontrollable, internal factors (the disability), but high performance to unstable causes.

Relevant to the relational demography framework, intergroup theory also proposes that attributions differ for in-group versus out-group members, consistent with the in-group bias predicted by social identity theory. Sometimes referred to as the ultimate attribution error (Pettigrew, 1979), successes of in-group members tend to be attributed to internal factors, and their failures to external factors. For out-group members, attributions show an opposite pattern, such that successes are attributed to external factors and failures are attributed to internal factors.

Research on attributions in social psychology has tended to corroborate these patterns. Predicted differential attributions for in-groups versus out-groups have been found (Beal, Ruscher, & Schnake, 2001). Gender effects on attributions also support predictions. A meta-analysis (Swim & Sanna, 1996) suggested that the more masculine a task is, the more likely it is that females' failures are attributed to stable causes such as low ability and task difficulty, while males' failures are more likely to be attributed to unstable factors such as low effort and bad luck.

Some empirical work in management has examined the role of attributions in appraisals. Heilman and Haynes (2005) used written scenarios of mixed
sex dyads completing a traditionally male task to manipulate clarity of contributions by varying task structure, information about individual contribution levels, and evidence of past performance. Consistent with predictions, when contributions and past performance were unclear, women were rated as less competent and influential and were viewed as being less likely to have filled the leadership role. When the task was structured in a manner that increased the clarity of contributions, gender effects on ratings were reduced. Greenhaus and Parasuraman (1993), studying race, found that relative to White managers, Black managers were less likely to have their work performance attributed to their effort and ability, and more likely to have their performance attributed to help from others. Thus, attributional patterns may be one of the processes through which ratings are affected.

Norm Violation (Backlash)

Eagly and Karau's (2002) and Heilman's (2001) stereotype fit models of gender bias further argue that evaluations of high-performing women in male sex-typed jobs are biased downward because such women are viewed as violating prescriptive aspects of the gender stereotype. Gender role stereotypes contain not only descriptive elements that describe how men and women are (e.g., men are agentic; women are communal), but also normative, prescriptive elements that specify what men and women should or ought to do (e.g., women should be communal; men should be agentic; Rudman & Glick, 2001). Rudman and Glick (1999) argue that because agency and communality are seen as opposing traits, a woman who succeeds in a male role may be viewed as agentic, but she is also likely to be viewed as not communal, and therefore as violating the prescriptive norm. Such individuals are more likely to be derogated and disliked. Status characteristics theory also predicts that norm violations by low-status group members result in derogation of the target. Prescriptive elements of the stereotype are predicted to be strong when the dominant group is dependent on the subordinate group and therefore has a strong interest in maintaining its own superiority and dominance. Negative reactions to violations are responses to a threat to the status quo (Ridgeway, 2001). Because a woman's role violation is a greater threat to the status quo than a man's is, there should be asymmetrical effects of role violation, such that women are penalized more for not being communal than men are for not being agentic (Rudman & Glick, 2001).

There is some evidence in support of this prediction. Heilman, Wallen, Fuchs, and Tamkins (2004) reported that successful women in a male sex-typed role were less liked by evaluators and more likely to be personally derogated than successful males. Being disliked had a negative impact on performance evaluations. They also found weaker evidence that successful men in female and neutral-typed jobs were disliked. In Rudman and Glick (1999, 2001) asymmetric effects were also reported. For women but not men, being
seen as agentic had a negative impact on perceptions of communality. Agentic females were rated as equally competent to but less socially skilled than agentic men were. These evaluations influenced perceived hireability ratings when job descriptions stressed interpersonal skills as well as competence. Rudman and Glick (2001) concluded that ironically including interpersonal dimensions as criteria for hiring and promotion may result in more and not less discrimination against women.

Double Standards and Extremity Effects

Bias is often thought of in terms of lower, less favorable evaluations. But another process involved in appraisal bias, concerns double standards for evaluating performance, which may not always result in lower evaluations. Stereotype fit and status characteristics theories argue that because group stereotypes carry expectations of competence, the level of performance needed for judging an individual as highly competent or capable will be higher for low-status than high-status group members (Eagly & Karau, 2002). Successful performance by a low-status individual (e.g., woman, person of color) will be assessed by a stricter standard than similar performance by a high-status person (Foschi, 2000). The shifting standards model (Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994) further proposes that judgments of minimum competence are also affected by double standards. For low-status groups, the level of performance needed for a judgment of minimum competence will be lower than it will be for high-status groups, again because of low expectations.

These arguments suggest that in general, high performance by a low-status or job-incongruent group member will not be rated as favorably as a similar performance by a high-status group member; and conversely, that poor performance by a low-status group member will be rated more favorably than the same performance by a high-status group member will be rated. Some studies in the management literature have reported this pattern of ratings. Gundersten, Tinsley, and Terpstra (1996) had senior business students play the role of a retail store manager appraising a subordinate’s performance based upon a performance log and a video. Female subordinates were given lower ratings than males were under high-performance conditions and higher ratings than males were under low-performance conditions. Similarly, Lynch and Finkelstein (2006) found that for poor performance there was a positive bias, but for high performance, a negative bias for disabled ratees.

However, research has sometimes reported an opposite pattern, such that ratings of stigmatized targets are polarized or extrematized in the direction of their performance (Fleming, Petty, & White, 2005). There are several explanations of this extremity effect. Expectancy violation theory argues that high performance by a low-status group member is rated more favorably than high performance by a high-status group member because high performance violates low expectations for the low-status person (Jussim, Coleman, & Lerch,
1987). Expectancy violation theory also predicts that among low performers, in-group members may receive lower evaluations than out-group members may because low performance by the in-group violates high expectations. In addition, social identity theory suggests that derogation of a negative in-group member can be one strategy to maintain the positive distinctiveness of the in-group (Tajfel, 1982). This "Black Sheep effect" predicts more negative ratings for in-group members than for out-group members at low levels of performance (Marques, Yzerbyt, & Leyens, 1988). Complexity-extremity theory (Linville, 1982) proposes that evaluations of out-group members are more extreme because raters have less contact with and thus a simpler cognitive schema of out-group members than of in-group members. Raters view the performance of an out-group member using fewer dimensions, leading to more extreme evaluations. Ambivalence-response amplification theory (Katz & Glass, 1979) argues that raters are ambivalent in their attitudes toward the stigmatized. These ambivalent feelings threaten the rater's sense of self, and extrematized responses occur as a way to resolve identity threat (D. L. Stone & Colella, 1996).

These theories have been variously discussed in models of bias to explain the positive bias sometimes found for ratings of disabled (Colella et al., 1997; Lynch & Finkelstein, 2006; D. L. Stone & Colella, 1996) and Black (E. F. Stone, D. L. Stone, & Dipboye, 1992) ratees, but little empirical work in management has directly examined predictions.

Shifting standards theory (Biernat & Manis, 1994) proposes that the response scale used by raters may determine if extremity effects occur. This model argues that on subjective (i.e., Likert) scales, which are often used for performance evaluations, raters will use within-category standards when judging ratee performance. So, for example, a woman's competence will be rated relative to the rater's (low) standards for women, while a man's competence will be rated relative to the rater's (high) standards for men. Thus, on subjective scales, the ratings of both high- and low-status groups can appear to be equally "good," as the scale conceals the double standards. This can produce the positive bias often seen for high-performing, low-status, or stigmatized ratees. The shifting standards model proposes that the effects of stereotypes (i.e., lower ratings for high performers; higher ratings for low performers) are more likely to be seen on objective scales or rankings (Biernat & Fuegen, 2001). Because these types of measures force raters to use the same scale to evaluate members of both groups, they are more likely to reveal the influence of stereotypes (Biernat & Kobrinowicz, 1997).

Biernat and her colleagues have conducted several examinations of shifting standards in performance appraisal contexts. In Biernat and Vescio (2002), participants played the role of coed softball team managers, judging the batting and fielding performance of male and female players. Although the information provided to participants equated athleticism for male and female players,
evidence of shifting standards was found on ratings of batting performance. Men were more likely to be rated as better batters on objective than subjective scales. However, no evidence of shifting standards was found on another dimension; the rating scale used did not moderate gender effects on ratings of fielding performance. Biernat et al. (1998) examined shifting standards for both race and gender in a sample of U.S. Army officers. As predicted, gender differences favoring males were found in rankings but not ratings of leader competence. However, there was no evidence for the use of different standards by race. Rather, both ratings and rankings favored Whites over non-Whites. The authors suggested that these results may have occurred due to the U.S. Army's explicit policy against using different standards based on race.

Research on bias in the management literature has not examined shifting standards directly. However, some results are suggestive. For example, Colella et al. (1998) found a negative bias only in rankings, not ratings, of disabled ratees, consistent with shifting standards theory.

**Emotion and Affect**

Dipboye (1985) noted that models of bias in the management literature emphasized cognitive processes while tending to ignore affective and emotional factors that might contribute to bias in appraisal. Although in general this is still true of models of bias, affect and emotion are hypothesized to play a role in the two major theoretical frameworks we have reviewed. We have seen that the stereotype fit model of gender bias and status characteristics theory propose negative affect toward those who violate status norms and threaten the dominant group's status. There is some evidence supporting this for gender. The relational demography approach suggests liking and positive affect as a mediator of similarity effects on judgments. There is also some evidence for this predicted role of affect. Tsui and O'Reilly (1989) reported that gender dissimilarity had a negative impact on liking as well as performance evaluations. Using structural equation modeling, Ferris, Judge, Rowland, and Fitzgibbons (1994) found that supervisor affect mediated the relationship of demographic (e.g., race, age, and function) similarity to supervisor ratings of subordinate performance.

Models of disability bias have given emotion a larger role, and theorists have expanded the consideration of emotional responses beyond positive and negative reactions and their effect on appraisals. D. L. Stone and Colella (1996) and Colella and D. L. Stone (2005) propose that disabilities evoke a variety of emotional responses in perceivers, ranging from mild discomfort to disgust or fear, or to pity and compassion, and that these emotions can have different effects on ratings. For example, pity and paternalism may result in a positive bias in ratings (Colella & D. L. Stone, 2005), while anxiety may have negative effects. However, we found no empirical studies of the role of emotions on appraisals for disabled ratees.
Recent work in social psychology has also examined the specific emotions associated with stereotypes of different groups. Fiske, Cuddy, Glick, and Xu (2002) proposed that stereotype content varies on two dimensions of warmth and competence. Groups considered high on both of these dimensions (e.g., one's in-group) are liked; groups seen as warm but not competent (e.g., elderly, disabled) receive pity; groups viewed as low warmth but high competence (e.g., Asian Americans) evoke envy; and groups viewed as low on both dimensions (e.g., poor people) evoke disgust. Cottrell and Neuberg (2005) argued that when social identities are salient, people will view out-groups in terms of the potential harm or benefit for one's in-group. The emotional response evoked by different out-groups is based on the type of perceived threat that the particular group poses to the in-group. An empirical study supported predictions, finding that out-groups seen as a threat to personal safety evoked fear; those seen as unable to reciprocate in exchange relationships due to an inability to respond evoked pity, and groups perceived as threatening values evoked disgust. These perceptions of threat and opportunity arise not only from the historical relations between the groups, but also current group relations (Cottrell & Neuberg, 2005).

These models move from consideration of a simple positive-negative reaction associated with a single in-group/out-group distinction to greater differentiation of emotions associated with specific groups, and also imply that emotional responses are inextricably linked to the content of stereotypes. Cottrell and Neuberg's (2005) theory of emotional responses as dependent on the threats posed by a particular out-group is consistent with status characteristics theory, which argues that the negative emotional reactions to some groups (norm violators) are due to perceived threats to the dominant group's power and status.

In summary, theory and research on processes presents a muddier picture given many theories, some with competing predictions. However, it seems likely that several processes are involved in bias. Evidence exists for the role of differential attributions and backlash against agentic women. For double and shifting standards, evidence exists but questions remain about the conditions under which effects will be seen. Most of the empirical work has concerned gender. Thus, one question concerns the extent to which these processes generalize to other categories. Extremity effects and differences in attribution patterns have also been reported for race (Black-White groups). Rudman and Glick (2001) propose that backlash against norm violators is unique to gender. They argue that communality (niceness) as a prescriptive norm is particularly strong for women because of men's dependence on women to fulfill important subordinate roles, which gives men a strong incentive to control women to maintain their status. When Black-White relations were also characterized by a similar close dependence between groups (e.g., in the pre-Civil War southern United States), communality was also prescribed for Blacks, and violators
more likely to be derogated or punished. However, societal changes that have enhanced the status of Blacks and decreased the dependency of Whites on Blacks have weakened the prescriptive stereotype for Blacks, while men continue to rely on women (Rudman & Glick, 2001). Therefore, prescriptive stereotypes for women remain strong.

However, Tsui, Xin, and Egan (1995) and Perry, Kulik, and Zhou (1999) suggested that violation of status norms may also contribute to age dissimilarity effects on evaluations. Since age is often used to infer wisdom and knowledge, an older subordinate may perceive a younger supervisor as lacking a legitimate basis of authority. One effect of this is that younger supervisors may view older subordinates as a threat to their higher status, leading to less liking and lower evaluations.

Evidence also exists for the predicted role of affect (e.g., liking) in similarity effects, but other predicted processes such as attributions and extremity effects have not been examined within the relational demography framework in management research.

**Contextual Factors Affecting Bias**

One of the unique contributions of the models of appraisal bias in management is their organizational focus. Although based on social cognition, the models move beyond a sole focus on cognitive processes to delineate contextual factors influencing bias (Kulik & Bainbridge, 2005). This organizational focus is important for not only understanding how and when bias occurs, but also identifying interventions that can reduce bias in appraisals.

The models of bias within the stereotype fit framework include contextual factors that affect the activation and use of stereotypes in making judgments. As the models are based on social cognitive theory and research, many of the contextual factors incorporated have also been derived from work in social cognition. For example, cognitive research has found that the amount of time available to decision makers, cognitive busyness, the amount, and quality of information available, and outside accountability can all influence the extent of stereotypic processing (Kulik & Bainbridge, 2005). These factors are included in several models of bias (e.g., Eagly & Karau, 2002; Heilman, 1983, 1995; Perry, 1997; Perry et al., 1994; Perry, Kulik, & Bourhis, 1996).

Beyond directly incorporating context variables found to be important in social cognitive research, models of appraisal bias delineate how these contextual variables can be influenced by organizational factors. For example, studies have found that rater accountability can decrease reliance on stereotypic processing (Perry et al., 1994). Models propose that rater accountability will increase, and bias decrease with the strength of the EEO culture, awareness of legislation (Colella et al., 1997; Powell & Butterfield, 2002; D. L. Stone & Colella, 1996), firm visibility, and the amount of external scrutiny received by the firm (Perry & Finkelstein, 1999). Both D. L. Stone and Colella (1996)
Bias in Performance Appraisal • 633

and Perry and Finkelstein (1999) propose that organizational technology will influence the extent of bias against disabled and older workers, respectively. Technology can influence the degree to which disabled employees are viewed as suitable for jobs by affecting the flexibility with which jobs can be performed (D. L. Stone & Colella, 1996). High-technology firms and jobs are likely to have a greater concentration of young workers than low-technology firms are, affecting the degree to which jobs are age typed and the salience of age in the setting (Perry & Finkelstein, 1999). The presence of internal labor markets has also been proposed to affect the salience of age, to the extent that older employees are in the minority in the incumbent populations of entry and lower level jobs. This may increase the activation of age stereotypes (Perry & Finkelstein, 1999). Heilman (1995) suggested that the presence of programs that target women as a group (e.g., mentoring, work arrangement, or affirmative action programs) will increase the salience of gender in the organization, and the extent of bias in appraisals and treatment.

Organizational culture, climate, and values are also proposed as influences on the salience of identity group memberships and the activation of stereotypes. For example, Perry and Finkelstein (1999) propose that age will be more salient in organizations that value age-associated traits such as flexibility, adaptability, and creativity. Ely and Meyerson (2000) and Cleveland, Vescio, and Barnes-Farrell (2005) argue that organizational cultures can be "gendered" in that informal and social patterns implicitly place a higher value on male identity and traits. Some examples of such gendered patterns are the use of time spent at work as a measure of organizational commitment and contributions, crisis oriented work patterns, and the portrayal of ideal workers as individualistic, assertive, and independent (Ely & Meyerson, 2000), as all of these value the male over the female identity. These patterns increase the salience of gender and contribute to the sex typing of jobs, increasing the probability of appraisal bias. Cox (1994) also proposed that the values of the dominant group are reflected in formal and informal work policies and practices, which he labeled institutional bias. This serves to disadvantage culturally different organizational members. D. L. Stone and Colella (1996) proposed that organizational values of competitiveness, individualism, and assertiveness may decrease the extent to which disabled workers are viewed as suitable for jobs, by implicitly valuing traits associated with "able" individuals.

Thus, models of appraisal bias have adapted and extended the results of social-cognitive research to develop rich and detailed predictions about how organizational factors may influence cognitive processing and bias. However, empirical research has not examined predictions regarding organizational variables. Instead research has examined if and how the contextual factors identified by social-cognitive work as influencing cognitive processing similarly influence the performance appraisal process. For example, for age, race, and gender, studies have shown that differences in ratings are less significant
when clear performance information is provided to raters (Finkelstein et al., 1995; Gaertner & Dovidio, 2000; Robbins & DeNisi, 1993; Weiss & Maurer, 2004). Cognitive busyness, the extent to which the rater is engaged in competing tasks, appears to make stereotyping more likely (Martell, 1996; Perry et al., 1996). So, rich theory but little evidence exists about how organizational factors influence cognitive processing and bias in appraisals.

Theoretical work on contextual and organizational variables influencing bias has been less apparent in the relational demography framework (Riordan et al., 2005). Tsui et al. (1992) argued that ratee tenure with a manager or group will moderate the effects of similarity on outcomes. They suggested that, with increasing time together, individuals have more opportunities to observe and collect accurate performance information, which should result in less reliance on stereotypes in making judgments. However, theories of intergroup relations argue that mere time together will be insufficient to improve cross-group relationships and decrease stereotyping (Dovidio, Gaertner, & Kawakami, 2003). The Contact Hypothesis (Allport, 1954) posits that intergroup interaction must occur under certain conditions: equal status, cooperation toward shared goals, opportunities for friendship, and support of authorities to decrease the tendency toward in-group/out-group categorization and the use of stereotypes (Allport, 1954; Dovidio et al., 2003). The proposal of Riordan et al. (2005) that interdependence may moderate similarity effects reflects this position. In addition, some work suggests that cooperative or collectivist cultures weaken similarity effects in groups (Chatman, Polzer, Barsade, & Neale, 1998). If as proposed, an emphasis on collectivist values decreases the salience of demographic categories as a basis for identification, collectivist values and culture may also moderate the effect of dyadic similarity on appraisals.

Tsui et al. (1995) proposed that work group heterogeneity moderates the effect of dyadic similarity. For example, for a female superior, they argued that the effect of subordinate similarity will be weaker when the work group is largely female, and stronger when the work group is largely male. When the work group is largely female, there are more opportunities to identify and interact with similar others, weakening attraction to particular subordinates. In contrast, when the work group is largely male, there may be greater attraction and identification with the few females in the group.

There has been little empirical investigation of these proposed contextual influences on similarity effects and performance evaluations. However, one longitudinal study found that similarity effects for White and male raters increased, rather than decreased over time (Biernat et al., 1998).

**Individual Differences Affecting Bias**

Rater individual differences are also included in models of bias. Theories of gender bias propose endorsement of traditional gender norms as a moderator of the backlash effect—negative reactions to women succeeding in
traditionally male roles (Eagly & Karau, 2002). Negative reactions are proposed to be greater for those who endorse traditional gender roles. Findings have partially supported this theory. Glick, Diebold, Bailey-Warner, and Zhu (1997) distinguished among several dimensions of sexism. Male endorsement of traditional gender norms that idealize women (benevolent sexism) did not result in lower overall evaluations of career women. However, views of women as inferior in ways that legitimize male social control (hostile sexism) predicted negative overall evaluations of career women for male raters. It may be that beliefs that women are inferior and should be socially controlled by men rather than the idealizing of women into certain traditionally feminine roles results in bias toward nontraditional women. Among female raters, ambivalent sexism (endorsement of both traditional gender norms that idealize women and views of women as inferior in ways that legitimize male control) predicted negative overall evaluations of career women.

Other models using the stereotype fit framework include individual differences as an influence on the activation and use of stereotypes. For example, Perry (1997) proposed rater authoritarianism and prejudice against older workers as affecting the extent to which age stereotypes will be activated and used in decision making. D. L. Stone and Colella (1996) and Colella et al. (1997) suggested that previous contact with the disabled should influence the use of stereotypes. Those with more previous contact are more likely to seek individuating information about the disabled, rather than to rely on stereotypes. In addition, personality variables such as self-esteem, ethnocentrism, and tolerance for ambiguity are predicted to influence attitudes toward disabled workers (D. L. Stone & Colella, 1996). These predictions have not received much empirical attention in the appraisal context. Perry et al. (1996) measured the extent of rater endorsement of age stereotypes and their impact on selection decisions. Those with higher stereotype endorsement evaluated age congruent job applicants more favorably than they did age incongruent applicants, with the greatest difference in evaluations on young-typed jobs.

Historically, most research on race bias has ignored individual differences (Roberson & Block, 2001). However, Brief and his colleagues (Brief, 1998; Brief & Hayes, 1997) have argued the importance of studying racism for understanding race discrimination in the workplace. They suggest that race bias and discrimination can be explained in part by modern racism (McConahay, 1983), a subtle form of bigotry which influences behavior only when actions can be attributed to a nonracial justification. Brief, Dietz, Cohen, Pugh, and Vaslow (2000) found that only when White participants were given a business justification to discriminate (the boss's belief that the race composition of the sales force should match that of the customer base), did modern racism predict selection bias against Black applicants. Aquino, Stewart, and Reed (2005) suggested social dominance orientation as a factor in appraisal bias. Those high in social dominance orientation believe in group hierarchies and
unequal social outcomes, and tend to be prejudiced toward a variety of groups (Dovidio & Hebl, 2005; Pratto, Sidanius, Stallworth, & Malle, 1994). Aquino et al. (2005) found that people high in social dominance orientation evaluated Black ratees more negatively than people with low social dominance orientation. This effect was stronger when the respondents believed that the Black ratee had been hired through an affirmative action process.

Although there has been little attention to individual differences in theory and research on relational demography (Riordan et al., 2005), some research in social psychology suggests individual differences that may moderate effects of similarity on appraisals. Branscombe, Wann, Noel, and Coleman (1993) and Biernat, Vescio, and Billings (1999) found that polarization of in-group ratings, including the tendency to derogate low-performing in-group members, was highest for those who identify strongly with their in-group. Biernat et al. (1999) proposed that high identification with the in-group activates high expectations for in-group members. When these are violated, negative evaluations result. These results suggest that rater level of in-group identification should influence the effect of similarity on evaluations such that similarity should have a stronger impact on the ratings of highly identified raters.

**Interventions to Reduce Appraisal Bias**

In 1980, Landy and Farr concluded that not enough was known concerning the dynamics underlying race and gender effects on performance appraisal to recommend interventions to eliminate biases. The past 25 years of research have increased our knowledge of the dynamics involved. Based on the research, we can conclude that giving raters more performance information, uninterrupted time to make appraisals, as well as increasing rater accountability are likely to decrease bias (Arthur & Doverspike, 2005).

In addition, several studies have examined the extent to which raters can reduce bias by using deliberate processing strategies. The effects of stereotypes on judgment are believed to be greatest when individuals rely on quick heuristic processing (Fiske & Neuberg, 1990), and use of a more deliberate, controlled strategy may decrease the extent to which stereotypes are used in judgments. Bauer and Baltes (2002) and Baltes and Parker (2000) used a structured free recall intervention to reduce rater reliance on heuristics. Raters were instructed to recall observed behaviors and use that information in making judgments. In Bauer and Baltes, this strategy decreased the impact of gender stereotypes on accuracy and level of performance ratings. In Baltes and Parker, free recall was also effective in reducing the performance cue effect, which occurs when performance expectations or past ratings of a ratee influence current ratings. Kulik et al. (2000) examined instructions for raters to suppress stereotypic thoughts. Previous research in social psychology had found that efforts to suppress stereotypes can backfire, making them more and not less likely to influence impressions at a later time (Macrae, Bodenhausen, & Milne, 1998).
Consistent with predictions, they found that raters who received instructions to suppress age-related thoughts and were cognitively busy during the rating task evaluated older job applicants most negatively.

Conclusions
By the 1990s many appraisal scholars had started to question the contributions of the psychological, cognitive approach to understanding performance evaluations (Arvey & Murphy, 1998). However, in the study of appraisal bias, cognitive models still have theoretical dominance and psychological theories have enhanced our understanding of bias (Kulik & Bainbridge, 2005). Our review finds support for both theoretical frameworks. The results of the review are shown in Table 13.1, which summarizes both theoretical and empirical research efforts so far. The left hand column shows, for each theoretical framework, the process theories, and moderators proposed as relevant for appraisal bias. The cells indicate, for each identity group, where empirical work has been conducted in a performance appraisal context.

For the stereotype fit framework, the central hypothesis (an interaction between job and ratee group membership) has been tested for each demographic variable with the exception of race. Although as we mentioned, status characteristics theory is relevant to race, no stereotype fit model of race bias has been proposed, perhaps reflecting the early skepticism of this model’s relevance for race (Ilgen & Youtz, 1987). This table shows that regarding processes, most empirical work has involved gender. For this variable, theory and research are most advanced, and they have benefited from research attention in other fields such as social psychology and sociology. Regarding the other demographic variables, it is clear that the processes involved in bias need more attention. This is particularly true for emotion and affect. Stereotype fit models have focused on the content of the stereotype as important for bias. Recent work linking stereotype content to emotions for various groups (Cottrell & Neuberg, 2005; Fiske et al., 2002) offers some directions for exploring the interplay of cognition and affect on bias in appraisals.

Looking at research on moderators, a few individual differences have been empirically examined, but many more have been proposed as relevant. However, the biggest gap between theory and research concerns organizational moderator variables. The table shows that the empirical study of situational moderators has exclusively involved testing the effects of contextual variables that influence cognitive processing. Research has investigated the amount/clarity of performance information, cognitive busyness, and the effects of rater strategies such as suppression and structured recall. However, an important contribution of management research to the study of bias has been to specify organizational factors that influence cognitive processing. Unfortunately, empirical research testing propositions about organizational variables and their effects on cognitive processing and appraisal bias is
<table>
<thead>
<tr>
<th>Theories &amp; Moderators</th>
<th>Stereotype Fit Framework Demographic Variables</th>
<th>Relational Demography Framework Demographic Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gender</td>
<td>Race</td>
</tr>
<tr>
<td>Performance Attributions</td>
<td>Empirical</td>
<td>Testing</td>
</tr>
<tr>
<td>Norm Violation</td>
<td>Empirical</td>
<td>Testing</td>
</tr>
<tr>
<td>Double/shifting Standards and Extremity effects</td>
<td>Empirical</td>
<td>Testing</td>
</tr>
</tbody>
</table>

**Moderating Variables**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational Variables Influencing Cognitive Processing</td>
<td>Empirical</td>
</tr>
</tbody>
</table>

Empirical Testing—indicates that there has been theoretical development followed by empirical testing of the theory or moderating variable for the demographic variable listed in the column heading.
virtually nonexistent. When organizational factors are mentioned, they seem to be invoked post hoc in an attempt to explain empirical findings. Kulik and Bainbridge (2005) noted that scholars fail to take advantage of opportunities to measure contextual variables in organizational settings, which would help explain the mixed results of research—why bias is seen in some but not all settings. Here we note that scholars also fail to take full advantage of the extensive theoretical work done in our field.

For the relational demography framework, the central hypothesis that rater–ratee similarity influences ratings has been tested for most of the demographic variables. However, this table shows that there has been little research on processes. Although the role of affect as a mediator has been explored in a few studies, other proposed processes such as extremity effects or differential attributions have not received attention in the management literature.

In contrast to the stereotype fit framework, very little work within the relational demography framework has examined potential moderators of similarity effects. Models of intergroup relations and the work of Cottrell and Neuberg (2005) point to the functional relationships between groups as a determinant of attitudes and stereotyping (Dovidio & Hebl, 2005; K. M. Thomas & Chrobot-Mason, 2005). Thus, intergroup relations within the organization should be important moderators of similarity effects. Further, Brief, Butz, and Deitch (2005) argued that perceived group relations outside the organization also influence relations in the workplace. Brief, Umpress, et al. (2005) found that the closer proximity with which White employees lived to Blacks, and the higher their perceptions of interethnic conflict in their community, the more negative their responses to a racially diverse workplace. Although appraisals were not studied, similarity effects may be stronger when intergroup relations in the external environment are poor.

Unlike gender, age, and disability, for race there has been less theoretical work in the appraisal literature on organizational factors that moderate the specific processes involved in appraisal bias. The diversity literature stresses the importance of organizational culture and climate for diversity as influencing outcomes for people of color, but the impact of these organizational factors on cognitive processing is not typically emphasized. For example, Cox (1994) argues that organizations pressuring assimilation of organizational members will result in poor outcomes for the culturally different. Might the extent of emphasis on assimilation also moderate the effect of similarity on appraisals? Dass and Parker (1999) proposed that some organizations are characterized by a climate "resistant" to diversity. In such organizations where diversity is viewed as a threat, perhaps out-group stereotypes are more salient and associated with greater negative affect. In addition, the research of Brief and his colleagues on the effects of racism and contextual factors on racial attitudes and discrimination is relevant to the appraisal context.
The theoretical frameworks have also proved useful in designing interventions to reduce bias in appraisals. Research suggests that raters be given information, time, and cognitive resources and also suggests that rater training might profitably teach strategies such as structured recall (Bauer & Baltes, 2002) prior to conducting appraisals. The study by Kulik et al. (2000) demonstrated the limitations of thought suppression instructions in a selection task. This suggests that organizations recommending stereotype suppression must also ensure that raters have sufficient cognitive resources when doing appraisals. However, information on organizational moderators of bias is also critical for designing interventions and settings to decrease bias in appraisal. For example, D. L. Stone and Colella (1996) proposed that when job technology constrains flexibility, disabled employees are less likely to be seen as suitable for a position, and bias in appraisal and selection are therefore more likely. Technology that enhances job flexibility should have the opposite effect. The implication is that technology can be used to design jobs that minimize bias. Others have proposed that attention to EEO legislation should decrease bias as it enhances rater accountability (Powell & Butterfield, 2000). On the other hand, if attention to EEO legislation results in greater salience of demographic categories to evaluators, this may increase the possibility of stereotyping and bias (Heilman, 1995).

From a relational demography perspective, research suggests that contexts which meet the conditions of the contact hypothesis (interdependence, cooperation) may decrease bias. In addition, a suggestion for rater training was found in a study by Crisp and Beck (2005), who evaluated an intervention rooted in the logic of social identity and social categorization theories. Their two experiments tested an intervention to reduce differentiation, the tendency of individuals to create cognitive distance between in-groups and out-groups, which in turn makes differences more salient. The manipulated intervention asked participants to think about the characteristics that are shared between the in-group and out-group before making judgments. This proved to reduce in-group favoritism on attitudes and resource allocation decisions, especially for low-identifying participants. The impact of this strategy on appraisal bias warrants investigation.

Limitations
This review was limited to a relatively narrow view of bias, defined as effects on performance ratings due to category membership. The two theoretical frameworks addressed in this chapter focus on rater effects—how similarity or job incongruence influences ratings through raters. In recent years, the concept of bias has been extended to include group differences in perceived fairness and justice in appraisals (Chen & DiTomaso, 1996), provision of feedback across group identities (Harber, 1998), and employee reactions to performance appraisal (Arvey & Murphy, 1998). In addition, some have
examined how rater bias may occur through ratee behavior. For example, Ferris and King (1992) proposed that age effects on performance appraisal operate through political behavior and ingratiation tactics used by ratees, which influence rater liking. Tsui et al. (1995) and Colella et al. (1997) have argued that the exchange relationship between superiors and subordinates also mediates the effect of similarity on outcomes. Similarity not only increases liking and in-group classification, it fosters a more favorable leader-member exchange. This can influence performance appraisals and performance level, as superiors give more support, information, and feedback in more favorable exchange relationships and subordinates respond with greater contributions and cooperation (Tsui et al., 1995). These broader issues of bias are important for understanding the effects of appraisals. For example, even if ratings themselves are not biased, appraisal feedback may be given differently, and may have differential impact across groups, which can affect outcomes such as advancement. To understand barriers to advancement fully, we need to know about bias in the broader appraisal system and feedback process as well as in the rating process.

In addition, we considered mostly research conducted in the United States and group identities important in the U.S. context. Bias is not unique to the United States, but the particular identities stigmatized or viewed as high or low status are likely to be culture specific. However, Schein (2001) found evidence of sex typing of the management role using Chinese, Japanese, British, and German samples. Role congruity (Rojahn & Willemsen, 1994) and similarity effects (Gibson, Zerbe, & Franken, 1993; Lindeman & Sundvik, 1995) have received some support through empirical testing outside of the United States, suggesting that the two theoretical frameworks are most likely generalizable.

Research also suggests that the factors salient for social identification differ by cultural context. Pelled and Xin (2000) found that age similarity had negative effects on relationship quality in their Mexican sample, but not in their U.S. sample. In addition, gender similarity had a stronger positive impact on trust in Mexico than it did in the United States, while it had a stronger positive impact on the leader-member exchange in the United States than it did in Mexico. Farh, Tsui, Xin, and Cheng (1998) examined relational demography predictions in the Chinese context and found positive effects of education similarity between rater and ratee; but not for age and gender similarity. They proposed that in China other factors such as guanxi (particularistic ties) may play a more dominant role in social identification than demographic variables. Other countries and cultures may demonstrate further differences.

Unique racial/ethnic compositions and historical contexts also provide a fertile ground for extending and generalizing theory such as in Western Europe where numerous excolonials, "guest workers," refugees, and other immigrants have settled in recent decades and face unique challenges in the work setting (Pettigrew, 1998). Similarly, research in countries with different
views on gender roles and beliefs about age may lend new insights into our understanding of bias.

In summary, understanding bias in appraisals is not only an issue of theoretical importance, but one of practical relevance. Where bias exists—where group identities matter—potential benefits of diversity for organizations cannot be fully realized and opportunities for individuals are limited. Management scholars have started to understand the dynamics of bias in appraisals. However, we do not yet know the extent to which bias limits the advancement of diverse workers, nor do we know how large a role appraisal bias plays in relation to the other mentioned barriers, such as lack of mentors or exclusion from networks. But models of bias argue that these barriers are likely to be related: the conditions that lead to appraisal bias are also hypothesized to lead to other forms of differential treatment that influence outcomes. For this reason, it is critical to give more research attention to the organizational conditions that influence the salience and use of stereotypes and similarity effects.

Acknowledgments
We wish to thank Elissa Perry for her comments and suggestions on an earlier draft of this chapter.

References


