AN EXPLORATION OF THE UNDERLYING MEANING OF JOB PERFORMANCE RATINGS FOR DIFFERENT ETHNIC GROUPS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy in the Graduate

School of The Ohio State University

By

Kathlyn Y. Wilson, M.A.

* * * * *

The Ohio State University

2003

Dissertation Committee:

Dr. Robert Billings, Adviser

Dr. Richard Jagacinski

Dr. Phyllis Panzano

Approved by

_____

Adviser

Department of Psychology

ABSTRACT

This study explored the underlying meaning of performance ratings to determine whether ratings may reflect different constructs across ethnic groups. Specifically, it was suggested that supervisors may emphasize a different set of factors across groups in arriving at an overall evaluation which would reflect different implicit theories of performance for different ethnic groups. Operationally, these differences were predicted to be reflected in two ways: 1) group differences in the interrelationships among performance ratings, and 2) differences across groups in the factors cited by supervisors in justifying their performance ratings of subordinates. Both hypotheses received partial support. Using a sample of bank staff, performance ratings were analyzed for potential group differences in terms of means and correlational relationships. Supervisors' written summaries of subordinate performance were content analyzed to identify the types of comments made across groups. The results are interpreted in light of the literature on group differences in performance ratings, and implications for future research and practice are discussed.

Dedicated to Yasmin and Zainab

ACKNOWLEDGEMENTS

I wish to thank my adviser, Robert Billings, for his support and patience which have made this dissertation possible. I want to thank my mother who provided the early inspiration to achieve academically. Thank you to Yasmin for the surprise notes, screen savers and words of encouragement.

VITA

October 21, 1958 ……..	Born – London, England
1982……	M.A. Counseling & Applied Services, Research & Evaluation Emphasis, The University of Tulsa
1985……...	M.A. Psychology, The Ohio State University
1988 – 1991……...	Consultant The Hay Group, London
1992 – 1998……...	Managing Consultant Independent Consultants, London
1992 – 1998……...	Associate Consultant Psychology at Work Institute of Psychiatry University of London
2003 – present	Instructor Department of Management & International Business College of Business Administration Florida International University

PUBLICATIONS

Wilson, K. Y. (1995). Appraisal: A Fair Assessment? CRE Connections, No. 4, April: Commission for Racial Equality, London, England.

Longenecker, C. O., Liverpool, P. R., and Wilson, K. Y. (1988). An assessment of manager/subordinate perceptions of performance appraisal effectiveness. Journal of Business and Psychology, 2:4, Summer.

FIELDS OF STUDY

Major Field: Psychology

Minor Field: International Business

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

"[P]redictors [are] subsidiary to the criterion…it is from the criterion that …predictors derive their significance. If the criterion changes, the predictor's validity is necessarily affected. If the predictors change, the criterion does not change for that reason..…research can be no better than the criteria used". (Nagle, 1953).

Studies of work performance have found significant differences between ethnic minority and non-minority performance. Findings are similar for ratings and objective performance measures (Ford, Schechtman, & Kraiger, 1986). This has been viewed as consistent with the finding that ethnic minorities also tend to score lower on some employment selection measures, the argument being that the lower performance ratings are validating performance on the predictor (F. L. Schmidt & Hunter, 1981).

Unlike the issue of test bias, the question of criterion bias has received very little research attention. There is little research investigating ethnic group differences. In contrast to the literature on gender effects in performance evaluation, for example, there is comparatively less research on race effects in appraisal. A psycINFO search using the key terms 'gender' and 'performance evaluation' returned 593 studies between 1985 and

2002. A similar search using the terms 'race' and 'performance evaluation' returned 138 studies during the same period. A search using a slight variation in key terms, 'gender' and 'work performance' and 'race' and 'work performance' returned 495 and 70 studies, respectively, during the same period. The focus of the few existing studies has been on identifying systematic biases in ratings as a function of rater and ratee race and determining the effect size (e. g. Landy & Farr, 1980) with no work on understanding or explaining the effect.

From a broad perspective, there are two possible explanations (which are not mutually exclusive) for these differences: 1) psychological and perceptual biases that might affect managers' ratings; and 2) real differences in performance that stem from differences in ability or differences in the sorts of experiences minorities may have at work (Ilgen & Youtz, 1990).

However, the extent to which we can accurately ascertain whether there are real differences in performance across groups or whether perceptual and other biases impact ratings, depends largely on our ability to accurately measure and assess job performance. Supervisory ratings of overall job performance serve as the primary criterion in validation research (Mount & Scullen, 2001; F. L. Schmidt, 2002). As such, these ratings constitute the basis for decisions regarding the validity of predictors for both research and practice..

Test fairness models assume an unbiased criterion, including Cleary's (1968) approach which is used in the EEOC's Uniform Guidelines on Employee Selection and Procedures (1978). Schmidt and Hunter (1974) and others (e.g. Guion, 1966) have warned that the concept of test fairness cannot be meaningfully examined without an

unbiased criterion. Yet, the primary focus of the test bias literature has been on predictor rather than criterion measurement (Schmitt & Noe, 1986).

Despite the importance of individual performance as a construct in psychology in general, and its critical role in determining predictor validity in industrial/organizational psychology, researchers have failed to apply the same degree of rigor and effort to developing criterion measures that they have used in developing predictors (Borman, 1991). Prior to 1990, performance as a construct has received very little research or theoretical attention (J. P. Campbell, McCloy, Oppler, & Sager, 1993). Studies that have sought to identify valid predictors of performance have been shockingly vague in their definitions of job performance. Jenkins' (1946) assertion that researchers are of the belief that "criteria are God given or to be found lying about" (p.23) generally still holds in current research. This "criterion problem" (Landy & Farr, 1983; Nagle, 1953) is not new. Austin and Villanova (1992) present a summary and analysis of the criterion problem dating from 1917 to 1992. Limitations in definition and operationalization of the criterion construct is a critical factor in the study of group differences as they arguably impact our ability to detect true performance differences across groups. Fundamental to explaining group differences in job performance ratings is an understanding of the criterion measures used.

The purpose of this research is to explore the underlying meaning of supervisory performance ratings, thereby gaining an understanding of what these ratings may represent and whether this may vary across ethnic groups.

This dissertation is organized as follows: Chapter 2 consists of four sections. The first reviews the existing empirical research on race differences in job performance. The

second section presents the primary methods used in identifying majority-minority group differences in performance evaluations. The third addresses the theoretical explanations for race differences in job performance. These are presented last as empirical research has generally preceded theoretical foundation. The focus of most research has been on identification of effects in the absence of explanatory frameworks. The final section presents the objectives of the present study and the specific hypotheses proposed. Chapter 3 presents the methods used, Chapter 4 the results, and Chapter 5 summarizes the findings and implications for future research.

CHAPTER 2

LITERATURE REVIEW

Empirical Research on Race in Appraisal

This section reviews the literature on minority group differences in work performance. The review begins with a chronological discussion of studies seeking to identify race effects in performance. These studies have been divided into two groups: those conducted between 1960 and 1986 and those conducted between 1987 and 2002. In 1985 and 1986, Kraiger and his colleagues conducted meta-analyses summarizing studies up to that point. It is useful to discuss the research in terms of work prior to and following these meta-analyses. Studies have been further organized according to setting – field versus laboratory research, and meta-analyses.

Early Research on Rater/Ratee Race (1960-1986)

The results of these early studies show a small, consistent race effect. Table 1 lists studies that have sought to identify race effects in ratings or other measures of work performance. Some studies have been included where identification of race effects in criterion performance was not the primary objective but criterion data were examined as a secondary aspect of the study. For example, test bias research (e. g. Baehr, Saunders,

Froemel, & Furcon, 1971; Farr, O'Leary, & Bartlett, 1971) where subgroup analyses were performed using criterion data and those findings were reported. Included are the research setting, whether or not group differences in ratings were found (Race Effect), the variance accounted for or size of the race effect if reported or applicable (the criteria referenced in this column are ratings unless otherwise noted); and the range of rate sample sizes. In some cases, variance accounted for or race effects were not reported. In other cases, multiple indices were reported (for varying criteria and for different samples). In this case, the data were not reported in the table as there was no overall statistic. Sample sizes are reported as ranges where more than one rate sample was used and data were analyzed separately.

Field Studies

Field studies generally show significant race effects. Some studies show clear race effects (e.g. Baehr et al., 1971; Bass & Turner, 1973; DeJung & Kaplan, 1962; Greenhaus & Gavin, 1972) while some have mixed findings (e.g. Farr et al., 1971; Hall & Hall, 1976). Only two studies found no race effects (Fox & Lefkowitz, 1974; Schmidt & Johnson, 1973).

In a sample of army recruits, DeJung and Kaplan (1962) found that ratees received higher ratings from members of their own race on a combat aptitude rating scale. Ratings were provided by fellow squad members. Their hypothesis that raters would give higher ratings to members of their own race was supported for minority recruits only. DeJung and Kaplan conclude that as there were very few black recruits in the squad, these men were rating their "closest buddies" (p. 373) possibly resulting in a preference and higher rating for these closer colleagues relative to other squad members.

Author(s)	Setting	Race Effect?	R^2/ Effect Size[a]	N
DeJung & Kaplan (1962)*	Field	Yes		32-370
Kirkpatrick, Ewen, Barrett, & Katzell (1968)*		Yes		71-535
Baehr, Saunders, Froemel & Furcon (1971)*	Field	Yes		60-188
Farr, O'Leary & Bartlett (1971)*	Field	Yes		18-322
Greenhaus & Gavin (1972)	Field	Yes		471
Toole, Gavin, Murdy & Sells (1972)*	Field	Yes		537
A. R. Bass & J. N.Turner (1973)*	Field	Yes		32-212
Beatty (1973)	Field	?[b]		44
Frank L. Schmidt & Johnson (1973)*	Field	No		93
Fox & Lefkowitz (1974)*	Field	No		67-100
Huck & Bray (1976)*	Field	Yes		35-241
Feild, Bayley, & Bayley (1977)	Field	Yes		9-33
Schmitt & Hill (1977)	Field	Yes		306
Cascio & Valenzi (1978)*	Field	Yes		911
Mobley (1982)*	Field	Yes	1%	1035
Thompson & Thompson (985)	Field	Yes	2%	150-233
Hamner, Kim, Baird, & Bigoness (1974)*	Lab	Yes	23%[c]	36[e]
Bigoness (1976)	Lab	Yes		60[e]
Hall & Hall (1976)	Lab	Yes No		290[e]

Table 1. Research on Rater/Ratee Race in Appraisal (1960-1986) by Research Setting (Continued)

Table 1. (Continued)

Author(s)	Setting	Race Effect?	R^2/ Effect Size[a]	N
Brugnoli, Campion, & Basen (1979)	Lab			56
Schmitt & Lappin (1980)	Lab	Yes		60[e]
Wendelken & Inn (1981)	Lab	Yes		551
Kraiger & Ford, 1985	Meta Analysis	Yes	3 to 5%	74 studies
Ford et al. (1986)	Meta Analysis	Yes	.16, .22[d]	53 samples

Note. *=Included in the 1985 meta-analysis by Kraiger and Ford. [a]These data are reported where available. [b]Only black supervisors were used; there was no comparison group. [c]Race and sex combined. [d]Objective measures and ratings, respectively. [e]Rater sample sizes.

In contrast, their majority counterparts were rating nearly all squad members. This operational effect would be "lost" for the majority sample in the lower average rating to the remaining squad members. The result is an apparent rating bias on the part of the minority group rater (DeJung & Kaplan, 1962). In a validation study of a test battery battery for police patrolmen, Baehr et al. (1971) found race differences in ratings and objective performance criteria. They also found that the best validation coefficients were for the black subgroup when the groups were analyzed separately; the poorest was when white weights were used in black equations and vice versa. The best cross-validation coefficients were obtained when weights based on a given racial group were used to predict scores for members of that group; i.e. when race-specific equations were used (Baehr et al., 1971). Greenhaus & Gavin (1972 found that majority ratees were rated higher on the three supervisory ratings used. In a study that examined criterion bias using

a sample of bank tellers, Bass and Turner (1973) found that white tellers were rated higher although the magnitude of the mean differences are small (less than half a scale point for performance ratings). In addition, correlates of ratings were different for black and white ratees. They found a broader set of correlates for white than for black tellers. More specifically, white tellers were evaluated based on a broader set of subjective data (e.g. alertness, cooperation, customer relations) and black tellers were evaluated based on a narrower set of ratings (e.g. quality of work). Related to the Bass and Turner (1973) finding, Beatty (1973) conducted one of the few studies aimed at examining the nature of the criteria used by employers in evaluating performance. In the study of a training program designed for the development of black supervisors, they found that black supervisors were not being rated on task-related factors or on factors related to the program content, but on other, more social, behaviors that they demonstrated at work. Factor analyses showed that a job performance variable was more heavily loaded on the rater's perception of black ratees' social behaviors than on their knowledge or task-related behaviors (Beatty, 1973).

Some studies had mixed findings. Whites received higher ratings in 13 of the 22 comparisons made in a series of validation studies using toll collectors, toll facility officers, correctional officers, and clerical staff (Farr et al., 1971). Toole et al. (1972) found that age moderated race effects. Whites tended to be rated higher in younger groups while there were no differences among older workers.

Cascio and Valenzi (1978) investigated the extent to which behaviorally anchored ratings of police performance were related to on-the-street objective indices of performance for minority and non-minority officers. They predicted that supervisory

ratings would be linearly predictable based on objective performance indices; and supervisory ratings would be more strongly related to objective indices for minority officers. This latter hypothesis was based on their interpretation of Bass and Turner's (1973) findings. They report that majority officers received significantly higher ratings; ratings were predictable based on objective indices; and objective indices were not more strongly related to ratings for minority officers, as both sets of correlations were statistically significant. Age and tenure were not rival hypotheses and accounted for only 1 percent of the explained variance (Cascio & Valenzi, 1978). Their findings may be a methodological artifact, however. In defining the set of objective measures, of the original 35 indices, 11 were eliminated due to low variances and low correlations with the BARS ratings! The remaining 24 were subjected to hierarchical multiple regression analyses for the two groups. Multiple R was not significant for either group and the set was further reduced to 8 variables. They acknowledge that the use of different objective indices may have produced different results (Cascio and Valenzi, 1978). It is not surprising, then, that ratings are predictable based on objective indices since those indices not correlating with ratings were dropped from the analyses. Also, the eight subjective criteria used (job knowledge, judgment, initiative, dependability, demeanor, attitude, relations with others, and communications) were highly intercorrelated (.84 to .91). Thus they used a linear composite. The high intercorrelations suggest halo and the probability that raters were using a single underlying dimension in appraising performance which, again, Cascio and Valenzi acknowledge. In addition, although they report no difference between the two groups in the relationship between ratings and objective measures, similar to Bass and Turner (1973), the pattern of intercorrelations is different between the

two groups. Specifically, intercorrelations between the BARS composite and the objective measures tend to be higher for the minority officers (black and Hispanic samples combined) and for the minority group all correlations were *negative*. Only one correlation (-.04) was negative for the majority sample and given its absolute value, it has no practical significance.

Mobley (1982) investigated adverse impact at a large supply distribution center. He tested the hypothesis of differences in rating variance as a function of evaluating same-race ratees. He found significant main effects for ratee race but not for rater race. Ratee race accounted for 1 percent of the variance in ratings. Contrary to previous findings, there was no evidence that raters evaluated their own subgroup higher than others.

Proportionally fewer studies report no race effects. Schmidt and Johnson (1973) found no race effects in peer ratings in a foreman training program. They point out that the fact that black trainees comprised 46 percent of the peer group and that the group had recently undergone "human relations" training may explain the lack of effect. The lack of effect given the size of the black sample would be consistent with Kraiger and Ford's (1985) later assertion regarding the proportion of minorities in a work group and the salience of race.

In developing a test battery for selection of entry-level employees in an electronics manufacturing organization, Fox and Lefkowitz (1974) found no evidence of criterion bias in a sample of black and white employees. There were no significant differences on the three criteria examined: 'efficiency scores' (percentage attained of a production standard), supervisory ratings, and supervisory rankings.

Laboratory Studies

Most of the lab studies reviewed identified significant race effects. However, the limitations to this paradigm are discussed below. In a simulated work-sampling task, Hamner et al. (1974) examined the effect of rater sex and race and applicant sex and race on overall task performance ratings. Subjects were undergraduate business students who were asked to rate student 'applicants' stocking grocery shelves. Subjects were assigned the role of grocery store managers. The task was to rate eight applicants who had applied for the job of stock worker in a grocery store. Applicants were videotaped viginettes performing a three-minute work-sampling task – removing large cans from an open case and placing them on grocery shelves. The performance was manipulated as high and low performing ratees. Subjects gave higher ratings to applicants of the same race. Hamner et al. also found a significant interaction between ratee-race and level of performance. Subjects differentiated more between high and low performing white than high and low performing black ratees. Specifically, black ratees were rated average at both levels of performance -- high performing blacks were rated only slightly higher than low performing blacks. High performing whites were rated significantly higher than low performing whites. Thus white ratees received significantly higher ratings overall at the same level of performance as black ratees. The result is that high performing whites and low performing blacks were rated more favorably compared to their counterparts. These performance differences accounted for 30 percent of the variance in ratings. Sex and race accounted for 23 percent additional variance in ratings. Hamner et al. (1974) conclude that "differential criterion bias is particularly prevalent for the high performing black applicant" (p.709). The findings also suggest that subjects tended to rate the performance

of white applicants more objectively than that of black ratees. These ratings would obviously be inappropriate for validation purposes since black and white ratings differ at the same level of performance (Hamner et al., 1974).

Using the same method as Hamner et al. (1974), Bigoness (1976) examined the effect of ratee sex, race and performance upon performance ratings conducted by a sample of white male undergraduates. Using the grocery-shelving task, they found no significant main effect based on race but significant interaction effects based on ratee race and performance. Among low performers, black applicants were rated more favorably than white applicants; there were no differences across race for higher performers; and standard deviations were highest for black males (See Table 2). Hamner et al.'s (1974) finding that white raters tended to rate black ratees lower was not replicated. Also, the tendency to favor high-performing white applicants was not replicated. One significant difference between the method used here and the Hamner et al. study is the fact that this sample of raters was comprised of only males whereas the earlier study included female raters. The relatively higher standard deviations in the black male ratings also suggest more variability in ratings even though performance levels were held constant across groups.

Hall and Hall (1976) extended the Hamner et al. (1974) and Bigoness (1976) studies by examining not just main race effects, but attempting to explore the role of stereotypes and rater characteristics in ratings for job incumbents rather than applicants. They hypothesized that: 1) there would be significant differences in performance evaluations of black versus white managerial incumbents displaying competent behavior.

Overall Task Performance	Black		White	
	Male	Female	Male	Female
Slow (low)				
M	7.12	7.18	6.38	6.28
SD	3.22	2.45	2.37	2.44
Fast (high)				
M	8.32	10.65	8.87	11.12
SD	3.03	2.58	2.53	2.84

Table 2. Means and standard deviations of ratings on overall task performance (Adapted from Bigoness, 1976; p.82).

That stereotypes, rather than performance data that defy conventional stereotypes, would drive performance ratings. 2) Perceptions and evaluations of black versus white incumbents would be related to demographic characteristics of the raters. Subjects were 290 predominantly white (222 white, 9 black, 59 unknown), male (212 male, 50 female, 28 unknown) undergraduates. A one-way analysis of variance yielded no significant differences across sex and race in mean scores on all dependent measures (three ratings). In terms of demographic characteristics of the rater, pearson correlations showed that race was significantly correlated with all performance ratings for the white male raters and not for the white female raters. Thus race of rater was a significant predictor for white males only. This is consistent with Hamner et al.'s (1974) finding that raters give higher ratings to members of their own race.

Schmitt and Lappin (1980) investigated the possibility that differences in validity coefficients across groups might be due to differences in criterion variance. They predicted that ratees would receive higher ratings from members of their own subgroup; rater confidence would be greatest when the ratee is a member of their own group; and the accuracy of ratings would be highest for ratees of the rater's subgroup. The stimulus

used was videotaped job samples of individuals shelving library books. Using a sample of 73 undergraduates, they found a significant main effect for race-of-rater (blacks were rated more positively) and a race-of-rater x race-of-ratee interaction. They found a correlation of .46 between reported confidence ratings and performance ratings confirming their hypothesis. They contend that rater confidence is greatest when raters are rating members of their own group; that when judging dissimilar people, raters are less confident and those ratings should display less variance as they are less likely to judge others at either extreme (very good or very poor). This finding is surprising in light of earlier studies (e.g. Bass & Turner, 1973) that found less variance and more halo in white ratings of white ratees. This is also not completely consistent with Hamner et al. (1974) who found that black and white raters differentiated between high and low performing whites (more variance in white ratings for both groups of raters) but high and low performing black ratees were rated average (less variance in black ratings). Correlations between ratings and actual performance were computed for each subgroup separately. Correlations of white raters' ratings with actual performance were higher than the corresponding correlations for black raters. Supporting their prediction, ratings of black and white subjects were more highly correlated with actual performance when they were rating members of their own racial group. Approximately 70 percent of the variance in rated performance was associated with actual performance (Schmitt & Lappin, 1980).

Meta-Analyses

In an often-cited study, Kraiger and Ford (1985) performed a meta-analysis of 49 published and unpublished studies reporting performance ratings of black and white

ratees. This included some of the studies reviewed here (see Table 1). They computed mean point-biserial correlations (corrected for unreliability in ratings) between ratee race and ratings for white and black raters. They found a small but consistent race effect across the studies. This effect was evident for both black and white raters. Corrected mean correlations were .183 (based on a sample of 74 studies and 17,159 ratees) for white and -.220 (based on a sample of 14 studies and 2,428 ratees) for black raters. Both tended to give significantly higher ratings to members of their own race

Moderator analyses for white raters showed that race effects were significantly higher in field (mean r_{pb}=.192) than in laboratory settings (mean r_{pb}=.037); effect size was inversely related to the proportion of minorities in the sample; and effect size was not influenced by type of rating (behavior/trait), rating purpose, or rater training. This finding for type of rating is significant since this suggests that rating formats such as behaviorally based ratings are as equally prone to race effects as trait ratings (Kraiger & Ford, 1985). This result is also consistent with Landy and Farr's (1980) conclusion that rating formats account for little variance in ratings. Even though Kraiger and Ford were not able to isolate race bias from true performance differences, there was some evidence that the results were due at least in part to rater bias. Firstly, raters evaluated same-race ratees higher than different-race ratees. As the two sets of raters evaluated many of the same ratees, one can conclude that the ratings were biased to some degree. Secondly, they found that the percentage of black individuals in the sample was inversely related to the size of the race effects (Kraiger & Ford, 1985)

Using a sub-set of these studies, Ford and his colleagues (Ford et al., 1986) conducted another meta-analysis. They were interested in examining comparative race

effects in objective and subjective evaluations of performance. Specifically, they used meta-analysis to 1) examine race effects for objective measures of performance and 2) compare the relative effect sizes for objective indices and subjective ratings of different criterion categories. The objective measures were comprised of three categories of performance: 1) a cognitive category (training and job knowledge tests); 2) absenteeism and tardiness, and 3) a performance indices category (direct performance: units produced, shortages; and indirect performance: accidents, customer complaints). The subjective ratings included overall ratings of effectiveness as well as specific ratings that matched the three criterion categories.

They found a small significant race effect for objective criteria. Effect sizes for subjective (overall ratings of effectiveness) and objective (all criterion categories combined) criteria are virtually identical. However, there were differences across and within criterion categories.

Specifically, there were significant differences in effect sizes *across* the three objective categories, with the largest effect size for cognitive criteria and smallest for absenteeism and tardiness (.336, .159, .112, respectively). Although Ford et al. (1986) report that differences in effect size across subjective ratings of the three criterion categories are nonsignificant, they actually report a significance value of $p<.05$. Thus, these results can be interpreted as significant. In the case of subjective ratings, there was no difference in supervisory ratings of the cognitive criteria and the performance indices categories (.23 and .22, respectively). Absenteeism and tardiness (.15) had the smallest correlation. Thus the pattern of results for the objective and subjective indices differ with

raters showing less discriminability on the performance and cognitive indices relative to objective performance on these same measures.

Within criterion categories (cognitive, absenteeism, and performance indices), cognitive measures (such as tests of job knowledge) had the largest significant differences relative to a matched set of subjective ratings. That is, there were larger race differences on objective measures of cognitive criteria than on ratings of the same cognitive criteria. Performance indicators also showed significant, albeit smaller, differences relative to a matched set of subjective ratings. Here, however, the direction of the effect was reversed – race effects were *larger* for subjective ratings of actual performance than for objective measures of actual performance. Thus majority group members were rated higher on performance indices than their level of performance on the objective indices.

The race effect in objective measures of actual performance was small (r=.16) suggesting that actual performance differences across groups are small. The effect size was the same for objective and subjective measures in the absenteeism and tardiness category (.11 and .15, respectively).

These moderator analyses suggest that majority members are rated higher on cognitive criteria, and equally high on performance indices than minorities; whereas, on objective indices, majority members are higher on cognitive measures but show no differences on performance indices. In addition, there is no difference in ratings or objective measures of absenteeism and tardiness.

This suggests that examining overall effect sizes for subjective and objective measures can be misleading as they obscure differences within different criterion

categories. In their review, Schmitt and Noe (1986) conclude that correlations between objective and subjective criteria are positive but generally low. Investigation of objective and subjective measures of different criteria shows race differences in objective indices and subjective ratings of the same criterion. Different measures of performance have different relationships "with the exogenous variable of race." (Ford et al., 1986; p.335).

Studies during this period suggest a significant, albeit small, race effect in performance measures. These effects are just as prevalent in the field as in laboratory studies.

Recent Research on Rater/Ratee Race (1987-2002)

While the earlier studies sought to identify race effects in performance measures across different contexts and criterion bias in validation samples, these later studies have sought to address some of the issues raised by the earlier research. The earlier work could be considered more exploratory in that research sought to answer the question: Are there race effects in performance measures? The more recent work, on the other hand, can be organized along three lines: 1) studies that respond to the call (Ford et al., 1986) for greater understanding of the constructs being measured by criterion measures including a focus on the job-relevant and job-irrelevant factors that underlie these measures. 2)Those studies that have sought to replicate or challenge some of Kraiger and his colleagues' (Kraiger & Ford, 1985) meta-analytic conclusions: that raters tend to give higher ratings to members of their own race (Lefkowitz & Battista, 1995; Mount, Hazucha, Holt & Sytsma, 1995; Prewett-Livingston et al., 1996; Sackett & DuBois, 1991; Waldman & Avolio, 1991); and that race effects decline as the percentage of minorities in the sample/workgroup increases (Pulakos, White, Oppler & Borman, 1989; Sackett, DuBois

& Noe, 1991). 3) Studies that focus on confirming a general race effect; that is, confirming differences in criterion performance across groups (Powell & Butterfield, 1997; Pulakos, White, Oppler, & Borman, 1989) and making the case for the use of a common regression line (Cleary, 1968) in predicting performance of minority and majority group members (Pulakos & Schmitt, 1996; Rotundo & Sackett, 1999). These studies are overwhelmingly field studies probably in response to some of the criticisms that have been levied regarding the external validity of laboratory studies. There has also been an attempt to increase generalizability of results by using large archival data sets. Research during this period is summarized in Table 3.

Understanding Underlying Constructs

Kraiger and Ford (1990) predicted that the pattern of relationships between supervisory ratings and objective performance indices would vary as a function of ratee race. They used meta-analytic techniques to aggregate studies that compared supervisory ratings to objective indices of job performance and job knowledge for black and white ratees. Their prediction (based on Cascio & Valenzi, 1978) of a stronger relationship between ratings and either objective criterion for black than white ratees was partially supported. They found a significant mean difference in correlations of ratings and objective performance data for black and white ratees (black ratee correlations were higher). For job knowledge, although the correlation between ratings and job knowledge data was slightly higher for black than for white ratees, the difference in mean correlations between the groups was not significant. They conclude that ratings of blacks by whites are more closely related to actual performance than are ratings of whites (Kraiger & Ford, 1990). They continue their call for meaningful research that

Author(s)	Setting	Race Effect	R^2/ Effect Size[ab]	N[d]
Pulakos, White, Oppler & Borman (1989)	Field/Project A	No	>1%	39,537 pairs
Greenhaus, Parasuraman, & Wormley (1990)	Field	Yes	2.5%, 5.5%[c]	828
Oppler, Campbell, Pulakos & Borman (1992)	Field/Project A	Yes		1823-3139
DuBois, Sackett, Zedeck, & Fogli (1993)	Field	Yes		1290
Mount, Hazucha, Holt, & Sytsma (1995)	Field	Yes		33-55476
Lefkowitz & Battista (1995)	Field	Yes		369
Prewett-Livingston, Feild, Veres & Lewis (1996)	Field	Yes		153
Pulakos & Schmitt (1996)	Field	Yes		464
Pulakos, Schmitt & Chan (1996)	Field	No		456
Powell & Butterfield (1997)	Field	Yes		300
Sackett & DuBois (1991)	Field/Archival	Yes		286-25,685
Sackett, DuBois & Noe (1991)	Field/Archival	Yes		814 wkgps
Waldman & Avolio (1991)	Field/Archival	Yes	1to8%	529-14403
Mount, Sytsman, Hazucha & Holt (1997)	Field/Archival	Yes		66-55706
Rotundo & Sackett (1999)	Field/Archival	Yes		229-17020
Stewart & Perlow (2001)	Lab	Yes		181
Kraiger & Ford (1990)	Meta Analysis	Yes		12 studies
Hauenstein, Sinclair, Robson, Quintella, & Donovan (2002)	Meta Analysis	Yes	d=.03	18 studies

Table 3. Research on Rater/Ratee Race in Appraisal (1987-2002) by Research Setting (Continued)

Table 3. (Continued)

Note. [a]These data are reported where available. [b]These are effect sizes or variance accounted for in ratings by group membership/race unless otherwise noted. [c]Ratings on relationships and task dimensions, respectively. [d]Rater and ratee sample sizes combined.

examines the extent to which raters differentially weight job-relevant and non-job-relevant factors when evaluating the performance of black and white ratees (Kraiger & Ford, 1990).

Meta-analytic studies of the relationship between ratings and objective performance measures showed that corrected correlations were low to moderate and ranged from .10 to .40 (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Heneman, 1986).

One study (Lefkowitz & Battista, 1995) directly investigated possible sources of criterion bias in supervisory ratings used as criteria for test validation. They employed five independent variables: employee ability, affect, rater's sex, rater's ethnicity, and expectancy effects (supervisor's knowledge of prior performance and supervisor's participation in hiring or promotion decision). Criterion ratings were: a *dimension composite* rating computed by averaging eight dimension scores; an *overall composite* based on the average of four global ratings; and an *average composite*, the average of the dimension and overall composite ratings. Lefkowitz and Battista (1995) report in their abstract that ethnicity was not associated with first-month ratings, but after five months, raters gave significantly higher ratings to same-ethnicity subordinates. However, this conclusion is not borne out in their results section or even in their own narrative report of their findings. They in fact state in the results section that "supervisors did not rate

subordinates of the same ethnicity higher than they did subordinates from different ethnic groups" (p.406). They present results from an ANOVA which show statistically significant main effects for employee race and supervisor race at one month and five months. Lefkowitz and Battista report that inspection of cell means and frequencies (not presented) shows that black supervisors were more stringent in their ratings of both ratee groups; black employees were rated lower by both raters; and employees were disproportionately assigned to supervisors of the same ethnicity as themselves at one month and moreso at five months.

A number of factors affect interpretation of these findings. Firstly, as half of the sample were black (46 percent), Kraiger and Ford's (1985) findings regarding the inverse relationship between race effects and the proportion of minorities in the workgroup would predict that race effects should be lower given the composition of this sample. Secondly, as black raters were more stringent than white raters and employees were disproportionately assigned to supervisors of their own race, ratings of black employees would be expected to be generally lower than their white counterparts. Lefkowitz and Battista concede that the lower ratings for black employees may be due in part to their disproportionate assignment to stringent black supervisors. They conclude, however, that despite the difficulty in interpretation given the between-subjects comparisons, their findings are consistent with conclusions by Sackett and DuBois (1991), and contradictory to those in other studies (Kraiger & Ford, 1985; Prewett-Livingston et al., 1996), that black raters do not rate black ratees higher than white ratees.

Interpretation of their results for liking was less problematic. Liking was the factor most highly correlated with the performance ratings at one month and five months,

independent of the effects of ability (Lefkowitz & Battista, 1995). Liking accounts for 30 and 33 percent of the variance in ratings at one month and five months, respectively. Liking was operationalized as supervisors' response on a five-point Likert scale to the item: "How much do you like him/her, regardless of their work performance?" Correlations are shown in Table 4.

	Rating		
Tenure	Dimension Composite	Overall Composite	Average Composite
1-month	.60** (235)	.60** (235)	.62** (235)
5-months	.68** (253)	.69** (253)	.70** (253)

Note. **p<.01. Figures in parentheses are the sample size.

Table 4. Correlations between performance ratings and liking for bank clerical staff at one- and five-months' tenure. (Adapted from Lefkowitz & Battista, 1995; p.400).

At five months, ethnicity was significantly correlated with liking (.13; p<.05). There was no such correlation at one month. In a sub-sample of employees with the same supervisor at one and five months, ethnicity was also significantly correlated with average ratings (.20; p<.01) at five months. In this sub-sample, ethnicity was significantly correlated with liking (.21; p<.01) -- the difference between liking and ethnicity correlations at one month and five months was statistically significant. Liking was significantly correlated with the average composite rating (.68; p<.01). Correlations for other performance ratings were not reported for the sub-sample. Given the substantial correlations between liking and performance ratings, it is conceivable that in addition to the observed correlation, ethnicity may have an indirect effect on ratings through liking.

There have been recent attempts to investigate race effects by partitioning the criterion space. Studies suggest a number of potential facets to the performance domain which can be grouped into two broad categories: 1) individual task performance, and 2) behaviors that create and maintain the social and organizational context that allows others to carry out their individual tasks (Murphy & Shiarella, 1997).

In a monte carlo simulation, Murphy and Shiarella (1997) showed that the level of a selection battery's validity varied substantially depending on how predictors were combined and how job performance was defined. They point out that large scale studies and meta-analyses (e. g. Hunter, 1986) have used measures (ratings) that confound individual task performance and broader social behaviors as supervisory ratings are probably affected by both. This would have implications for ability→performance validities in particular (per Schmidt & Hunter's model – e.g. Schmidt, Hunter & Outerbridge, 1986), and validities in general. Specfically, the "validity of selection tests for predicting complex criteria may show considerably less generalizability than current meta-analysis of univariate validities would suggest" (Murphy & Shiarella, 1997; p.823).

Borman and Motowidlo (1997) make the distinction between task and contextual performance. Task performance refers to task proficiency and reflects the individual's job knowledge, skills, abilities, experience and training (Hattrup, O'Connell, & Wingate, 1998). Examples include operating a machine, producing a written document and performing a surgical operation (Mototwidlo, Borman & Schmitt, 1997). Contextual performance, on the other hand, "involves behaviors that support the social, organizational, and psychological environment in which task behaviors are performed" (LePine, Hanson, Borman, & Motowidlo, 2000). Examples include such behaviors as

helping, cooperation, and volunteering. These reflect the individual's motivation, affect, and interpersonal orientation (Hattrup et al., 1998).

Preliminary research suggests that supervisors equally consider task and contextual performance in making performance judgments; i.e., that both factors contribute independently to overall performance ratings (Borman & Motowidlo, 1997; S.J. Motowidlo & Van Scotter, 1994; Organ, 1997). Further, the kinds of knowledge, skills, work habits, and traits associated with task performance are different from those associated with contextual performance (S. J. Motowidlo, Borman, & Schmit, 1997). Specifically, task and contextual performance have been shown to have different predictors (Hattrup et al., 1998; S.J. Motowidlo & Van Scotter, 1994). Task performance is predicted by ability whereas contextual performance is predicted by personality variables (Hattrup et al., 1998).

Johnson (2001) examined the relative importance of task and contextual elements of performance to supervisory ratings of overall performance. He found that job-specific task proficiency (one dimension of task performance) was the most important dimension in two of the eight job families studied. Non job-specific task proficiency was most important for the remaining six job families. These are tasks not central to any particular job and necessary in most jobs; for example, plan, organize, coordinate, and decision-making (Johnson, 2001). Further, in all of the job families, contextual dimensions made a unique contribution to overall evaluation. Thus although task performance dimensions were the most important for each job family, the other dimension on which supervisors placed most weight in overall performance were contextual dimensions (Johnson, 2001).

A number of factors might influence the relative importance of task and contextual performance (Johnson, 2001) such as the organization's culture and characteristics of the ratee. Gender stereotypes, for example, have been shown to influence appraisals (Heilman, Block, Martell & Simon, 1989). Thus it is conceivable that stereotypes may drive the relative importance given to each; suggesting that the relative emphasis may vary for minorities compared to non-minorities.

There is some evidence of ratee race effects in task performance (Hattrup, Rock, & Scalia, 1997; Hauenstein, Sinclair, Robson, Quintella, & Donovan, 2002) but less clarity regarding the impact of rate race for contextual performance. Based on existing research, Hattrup et al. (1997) postulated that white ratees would be rated higher on task performance relative to black ratees and there would be no difference between the two groups on contextual performance. They suggested that weighting of contextual factors more than task factors may result in less adverse impact; while weighting task factors more heavily would increase the weight placed on cognitive ability and thereby result in more adverse impact in the hiring process (Hattrup et al., 1997; Hattrup et al., 1998). The latter proposition is supported by Ford et al. (1986). There were similar race effects for ratings of both performance and cognitive criterion categories. Thus raters did not distinguish between those performance indices even where there were no differences between groups in objective performance measures of the same criteria. This suggests that the relationship between cognitive criteria and task performance in raters' implicit models, may result in lower ratings for minorities on task factors irrespective of the objectively measured relationship between cognitive criteria and task performance.

Hauenstein et al. (2002) performed a direct test of Hattrup et al.'s (1997) proposition using a meta-analysis of studies between 1970 and 2001. Using 18 studies and a total of 122 effect sizes, k (the number of effect sizes contributing to each effect size estimate) ranged from three to 45. Task performance was divided into five subdimensions: job knowledge, productivity/quantity, accuracy/quality, customer service, and general. Contextual performance categories were conscientiousness and other. This meta-analysis only partially supported Hattrup et al.'s (1997) findings. There were group differences on both task and contextual performance dimensions. Specifically, black ratees were rated lower on task dimensions (job knowledge, productivity/quality, accuracy/quality, customer service skills, and management/administrative skills) as well as some sub-dimensions of contextual performance (interpersonal skills, effort/initiative/enthusiasm). There were no differences on the conscientiousness sub-dimension (Hauenstein et al., 2002). This finding for contextual performance was in contradiction to Hattrup et al's (1997) prediction. However, this result would be consistent with research, which shows that supervisors consider social behaviors in rating minority group members (Beatty, 1973). Such behaviors would fall under the rubric of "contextual" performance.

Challenging Meta-Analytic Conclusions

Sackett and his colleagues (Sackett & DuBois, 1991; Sackett, DuBois & Noe, 1991) conducted a series of studies to challenge Kraiger and Ford's (1985; 1986) meta-analytic findings regarding rater-ratee race effects in performance evaluations. They used a large civilian data base, data gathered by the U.S. Employment Service (USES) from 1972 to 1987; a large-scale military study (Project A data analysed by Pulakos et al.,

1989); and a subset of Kraiger and Fords's (1985) meta-analysis (studies including black raters). The civilian data base included performance data for 174 jobs representing 2,876 companies. Individuals in the same job within the same company were considered a "workgroup." Ratings analyzed were averaged across multiple dimensions and represented overall performance. The military data was collected as part of Project A. A total of 6,377 supervisors and 8,174 peers evaluated first-term soldiers representing 19 military occupational specialities. The study compared Kraiger and Ford's findings with these two data sets. They found that black ratees received lower ratings from both black and white raters although the magnitude of difference was substantially larger with white raters. Reanalysis of these data excluding peer ratings resulted in small effect sizes in the military sample that were smaller than those resulting from a similar reanalysis of Kraiger and Ford's data.

Sackett, DuBois and Noe (1991) used the USES data base to examine the effect of tokenism on performance evaluation; i.e. that the proportion of women/minorities in a workgroup impacts performance ratings. They found that the proportion of black ratees in the group does not affect ratings. This finding may be due to their definition of "workgroup." Individuals in the same job in the same organization are not necessarily in the same workgroup in the traditional sense of the term. A more accurate conclusion would be: the proportion of black ratees in the same job in the same organization does not affect ratings. A different pattern of differences is obtained when women are tokens than when minorities are tokens (Sackett, DuBois & Noe, 1991). They speculate that the true differences between black and white ratees is large and the true performance difference between men and women small.

More recently, Rotundo and Sackett (1999) examined the issue of the potential impact of criterion bias on predictor validity. They attempted to decrease bias in criterion ratings by using same-race raters. They made between- and within-subjects comparisons of ratings of black and white ratees by same-race and majority group supervisors. They found that for the between-subjects sample, black ratees received lower ratings from black and white raters although the magnitude of the difference between the two raters was different. White raters rated black ratees lower than black raters. In the within-subjects sample, black ratees received significantly lower ratings than white ratees only from white raters, and not same-race raters. Also, validity coefficients for black ratees in the within-subjects sample were not significantly different from zero. Thus validities were significant only when black ratees were being rated by majority group raters.

Replicating Race Effects

In an attempt to isolate race (and gender) bias through repeated measures analyses, Pulakos et al. (1989) found no significant effects in ratings of Army personnel. They conclude that one explanation for this finding may be the large percentage of minority service members (in different jobs in the same organization); consistent with Kraiger and Ford's (1985) finding regarding the saliency of race.

Greenhaus et al. (1990) examined relationships among race, organizational experiences, job performance evaluations, and career outcomes for black and white managers. Using a sample of predominantly white raters (93 percent), they found significant race effects for the two job performance dimensions studied: relationships (2.5 percent of explained variance) and task (5.5 percent of explained variance). Explained variance is similar to to Kraiger and Ford's 3.7 percent. The variable, race, also

explained more variance in job performance evaluations than organizational experiences such as job discretion, acceptance, and supervisory support. Waldman and Avolio (1991) found that race accounted for up to 8 percent of the variance in ratings across different occupational groups; ranging from 1 percent in service and health care jobs to 8 percent in technical plant operations.

Summary of Existing Research on Race in Appraisal

Despite the inconsistent findings over the past 40 years, some tentative conclusions can be drawn from this research: both minority and majority group members give higher ratings to members of their own race (Kraiger & Ford, 1985; Mount, Sytsman, Hazucha, & Holt, 1997; Prewett-Livingston, Feild, Veres, & Lewis, 1996); performance ratings are affected by the composition of the workgroup – specifically, race effects decline as the percentage of minorities in a workgroup increases (Kanter, 1977; Kraiger & Ford, 1985; Prewett-Livingston et al., 1996; Pulakos et al., 1989); race contributes to variance accounted for in performance ratings although the explanation for this finding is unclear (Ford et al., 1986; Kraiger & Ford, 1985; Lefkowitz & Batista, 1995; Prewett-Livingston et al., 1996; Waldman & Avolio, 1991); studies have shown race to account for one to eight percent of the variance in performance ratings (e.g. Waldman & Avolio, 1991); and the race effect in actual job performance is small (Ford et al, 1986). This latter finding is particularly significant. As Ford et al. (1986) point out, it suggests that aptitude and job knowledge tests may measure some construct correlated with ethnicity but unrelated to actual job performance. Initial studies of group differences in task and contextual elements of performance show inconsistent results. There have been group differences on task performance (Hattrup et al, 1998) and group

differences on both task performance and sub-dimensions of contextual performance (Hauenstein et al, 2002). These findings, together, suggest the need for a greater understanding of the constructs being measured by criterion measures and, in particular, the constructs underlying subjective ratings.

Explaining or understanding group differences in job performance has not been the focus of past research. Further, the research methods used to identify race effects have clear limitations which may contribute to the inconsistency in findings. To begin with, the use of laboratory studies limits generalizability of findings to real world situations in several ways. One limitation of laboratory paradigms is the simple nature of the simulated job typically used. For example, stocking grocery shelves within a three-minute time frame (Hamner et al., 1974; and Bigoness, 1976); videotape of subjects shelving library books (Schmitt & Lappin, 1980); and rating paper people (Hall & Hall, 1976). Real world appraisals (ratings) are ostensibly made with as much supporting information as possible. With a short, simple task or paper people paradigm, the amount of stimulus information available to the rater is limited to a few salient characteristics presented at one point in time (Wendelken & Inn, 1981). Laboratory paradigms also deal with information that is complete and immediately available (stimulus-based) while performance judgments are typically memory-based (Feldman, 1981). Finally, in performance appraisal situations, raters may be trained to reduce rating errors (e.g. Schmidt & Johnson 1973) a factor which is not incorporated into laboratory approaches.

The more recent studies seem to confound rather than clarify our understanding of performance differences. In an attempt to increase generalizability of findings, the more recent literature is dominated by studies using large archival data sets of primarily

military (and some civilian) samples. Several studies have analyzed the same data set used in previous work (e.g. Pulakos et at., 1989; Sackett & DuBois, 1991). There are a number of limitations to this type of analysis: Firstly, the use of large archival data sets highlights the continued emphasis on identifying race effects rather than understanding the process by which these differences occur. Secondly, these studies use composite data gathered for over a decade. The length of time of data collection is a concern since appraisal ratings made in 1972, for example, may be qualitatively different from those made in 1987 (e.g. Sackett et al., 1991). The type of appraisal instruments developed over that 15-year period, for example, may have changed -- a point in case being the shift from graphic rating scales to behaviourally anchored ratings scales as a function of research during the 1980s. To collapse these data render them virtually meaningless. Thirdly, and most importantly, archival studies use data sets with different types of appraisal data gathered across a number of different jobs for different purposes; for example, Sackett and DuBois (1991) and Sackett et al.'s (1991) use of U.S. Employment Services (USES) data gathered between 1972 and 1987. The data set contained job performance measures for 174 jobs in 2,876 companies. Research conducted with such composite data lends itself to different types of conclusions and generalizations from research conducted in one organizational setting. In research involving one or perhaps two organizations, data are collected for more clearly delineated purposes, and more information is available about the performance measures used, the rater, the ratee, the organization, and the context in which the appraisal/performance measurement is made. For example, appraisal purpose can influence the accuracy of ratings. Studies have shown more differences across groups in ratings made for salary decisions compared to

ratings made for research purposes (Kirkpatrick et al., 1968; Wollowick, Greenwood, & McNamara, 1969). The lack of statistical control of variables such as tenure, job, organizational level, and functional area can mask important relationships (Lefkowitz, 1994). Ford et al's (1986) findings regarding race differences in objective and subjective measures of the same criterion is particularly relevant here. There were similar race effects for overall objective and overall subjective criterion measures (.209 and .204, respectively). However, there were differences in objective and subjective measures within a criterion category. For example, there was a larger race effect for subjective measures of performance indices (units produced, shortages, accidents, and customer complaints) compared to objective measures of the same performance indices (.221 and .159, respectively). Archival studies generally use overall ratings and/or summary ratings comprised of the sum of dimensional ratings (e.g. Sackett, DuBois & Noe, 1991; Waldman & Avolio, 1991) which may obscure relationships among the underlying individual criteria. Large-scale archival studies tell us more about the outcome than the process of performance measurement. Our need to generalize findings in research for purposes of economy in practice sometimes overrides the benefits of conducting organization-specific research that would provide a better understanding of relevant constructs.

There is a clear need for research that moves beyond identification of race effects to understanding the underlying processes involved in rating different ethnic groups and identifying the exact nature of the performance differences across groups. The methodological approaches that have been used in exploring race differences in

performance as well as the description of a potentially useful approach for understanding the constructs underlying performance ratings is the subject of the next section.

Methods of Identifying Rater/Ratee Race Effects

As most studies in the literature reviewed have identified race effects, the real issue becomes whether these effects are due to criterion contamination or whether they reflect real performance differences; and in the case of either, why? The extent to which this is addressed varies across studies. As stated earlier, the focus of existing studies has been on explaining variance rather than explaining the process. Researchers have generally used one (or a combination) of three approaches in the attempt to identify subgroup bias in performance measures: the Total Association (TA) approach, the Direct Effects (DE) approach, and the Differential Constructs (DC) approach (Oppler, Campbell, Pulakos & Borman, 1992).

Total Association Approach

This approach is used to determine the total amount of criterion variance accounted for by subgroup membership (Oppler et al, 1992). Several studies employed this approach. Most of these were field studies. DeJung and Kaplan (1962), for example, investigated the relationship between rater and ratee race and peer ratings on a combat aptitude rating scale. They predicted that: 1) ratees (army recruits) would receive higher ratings from members of their own race; and 2) raters would give higher ratings to men of their own race. Rater agreement correlations were computed across race. The hypothesis that ratees would receive higher ratings from members of their own race was supported for the four ratee samples used. The actual correlation coefficients were not reported in

the study. Differences between correlated means were tested using one-tailed critical ratios. None of the F ratios were significant. The within ratee covariation of majority and black ratings was examined using pearson moment correlations computed for the four sample groups studied (regular army versus inductee recruits for each race). The average rating for each ratee based on ratings by white and black raters correlated .52 and .52 for the two white samples, and .42 and .47 for the two black samples (p. 372). Their second hypothesis was supported for black raters but not for majority raters. The most significant of the total association studies is Kraiger and Ford's (1985) meta-analysis.

Total association studies suggest there is a small relationship between ratee race and performance ratings; moderated by race of the rater. However, this approach cannot determine whether the effects are criterion relevant or due to criterion contamination – whether they represent bias or true performance differences. Total association between two variables has the following possible components:

1. causes common to both
2. correlations among causes common to both
3. indirect causal effects of one on the other
4. direct causal effects of one on the other (Oppler, Campbell, Pulakos, & Borman, 1992)

Estimates of total association do not differentiate among these. Also, as estimates of total association are correlations, this approach does not provide any information about the process by which bias occurs (Oppler et al, 1992). As Kraiger and Ford (1985) unwittingly point out in discussing the objective of their meta-analysis, this method does not directly isolate the issue of racial bias in evaluations, but provides a necessary first

step by answering the question whether race is related to performance evaluation under certain conditions.

Direct Effects Approach

The attempt here is to isolate the effects of subgroup membership not mediated by true performance differences. This is done through experimental manipulation of performance in the laboratory or through statistical control in the field. Effects would suggest group-related criterion contamination (Oppler et al, 1992). For example, Bass and Turner (1973) examined supervisory ratings and measures of objective performance for black and white ratees. They were interested in identifying differences in performance measures as well as examining the extent to which ratings were biased as a function of ratee race. To determine the latter, they compared means across groups and statistically controlled for the effects of age and job tenure. They found that white ratees were higher on mean criterion scores. As age and tenure were correlated with various measures to some extent, they computed partial correlations between race and criterion measures where age and tenure were partialled out. For the full-time sample, race was significantly correlated with four measures before the partialling (attendance, number of shortages, number of overages, and customer relations) and two after (attendance and number of shortages). They found a similar pattern in the part-time sample. Race was significantly correlated with seven (supervisory ratings: customer relations, quality of work, alertness, cooperation, and overall effectiveness; objective criteria: adjusted salary increase, and number of shortages) performance measures and four (quality of work, overall effectiveness, adjusted salary increase, and number of shortages) were still significant after partialling. To investigate whether these remaining differences might

represent unfair discrimination, they computed correlations between ratings and objective measures separately for the black and white samples. For black tellers, only one rating, quality of work, is related to the objective criteria (adjusted salary increase, number of shortages, and number of overages (Bass & Turner, 1973)); whereas, for white tellers, five of the six ratings were related to salary increase and three of the six to number of shortages. Essentially, fewer ratings are correlated with objective measures for black versus white tellers, but the correlations for black tellers are higher showing a stronger relationship than for white tellers. A similar pattern holds for part-time tellers.

Even with the presence of race effects, studies using this approach would have to meet three assumptions before concluding that these effects represent real differences rather than criterion contamination (Oppler et al., 1992), p.203:

1. the covariates used to account for the effect must represent a relevant aspect of performance
2. the covariate itself must not suffer from subgroup contamination
3. there are no other factors contributing to the relevant variance that are also related to race; otherwise it would be based on an incomplete set of covariates

Not surprisingly, covariates in existing studies do not meet these assumptions.

Differential Constructs Approach

The differential constructs method represents a potentially useful approach to understanding group differences. Group differences is an established method of examining construct validity (Cronbach & Meehl, 1955). This approach investigates the extent to which the construct validity of ratings differs according to the race of the rater or the ratee. Racial bias in ratings can be "inferred when ratings hold different meanings

for different racial subgroups" (Kraiger & Ford, 1990; p. 269). Relevant data are correlations between ratings and other variables. The extent to which these correlations vary in size or pattern across different rater-ratee subgroup combinations indicates that the psychological meaning of the ratings depends on the subgroup of persons being rated, the subgroup of persons providing the ratings, or both (Oppler et al., 1992).

Several studies provide evidence for possible differences in the meaning of the criterion measure across groups. In an early study that sought to identify bias in criterion measures, Bass and Turner (1973) correlated supervisory ratings on six performance factors, including overall effectiveness, with four 'objective' criterion measures for a sample of black and white bank tellers. They report nonsignificant mean differences in these criterion measures once the effects of age and job tenure were partialled out. However, their results show, for the full-time sample, that with age and tenure partialled out, there were still small but significant correlations between race and two objective criterion measures: *attendance* (.17, $p<.05$), and *number of shortages* (-.24, $p<.01$). For the part-time sample, there were small significant correlations between race and supervisors' ratings of *quality of work* (.21, $p<.01$), and *overall effectiveness* (.17, $p<.05$); and two objective criterion measures: *adjusted salary increase* (.17, $p<.05$), and *number of shortages* (-.19, $p<.05$).

Inspection of the intercorrelations among the different criterion measures also reveals a different pattern of results for the two groups which were not highlighted in the study. Specifically, intercorrelations among supervisory ratings on the six performance factors were different for the two groups. There was more intercorrelation among supervisory ratings for whites than for blacks. While the intercorrelations were generally

high, they were higher for the white tellers. For example, for full-time tellers, *overall effectiveness* was highly correlated with *customer relations* (.79) and *alertness* (.84) for black tellers while it was highly correlated with *quality of work* (.74), *alertness* (.78), and *cooperation* (.75) for white tellers. Similarly, for part-time tellers, *overall effectiveness* was highly correlated with four other supervisory ratings for white tellers compared to two other ratings for black tellers. This suggests less dimensionality in the white ratings, perhaps more halo, and possibly a different factor structure for the two groups.

For both full-time and part-time tellers, there were more significant intercorrelations between ratings and objective criteria for white tellers than for black tellers suggesting possible criterion bias (see Figures 1 and 2). Correlations for black tellers were higher than those for whites, however, there were fewer significant correlations relative to the white sample. For full-time black tellers, only one rating, *quality of work*, was significantly correlated with objective measures; specifically, *number of shortages* (-.55, $p<.01$), *number of overages* (-.51, $p<01$), and *adjusted salary increase* (.37, $p<05$). For white tellers, there were more significant correlations among supervisory ratings and objective criterion measures. Five of the six performance ratings, *customer relations, quality of work, alertness, cooperation* and *overall effectiveness,* were significantly correlated with *adjusted salary increase*; and three ratings, *quality of work, alertness,* and *overall effectiveness* were correlated with *number of shortages*.

Also, for full-time tellers *overall effectiveness* was not significantly correlated with any of the objective criterion measures for the black sample, whereas it was correlated with *adjusted salary increase* and *number of shortages* for white tellers. Thus, *overall effectiveness* was not related to *salary increase* for either full-time or part-time

black tellers whereas it was for white tellers in both samples. In addition, salary increase was significantly correlated with attendance (.51; $p<.01$) for black tellers while there was no corresponding relationship for white tellers. This suggests that supervisors may consider the less objective performance factors when making salary decisions for whites (Bass & Turner, 1973).

This study found that fewer supervisory ratings predict objective measures in the black sample relative to the white group. It appears, then, that there may be less differentiation in supervisors' ratings of black tellers. That is, although attitudinal and motivational factors (e.g., cooperation, customer relations, alertness) are taken into account in rating white tellers, it is less clear on what supervisors are basing their evaluations of black tellers. This clearly suggests that there may be differences between the two groups in the underlying meaning of the criterion measures used. The authors conclude that selection tests showing differential validity may be reflecting differences in the nature and meaning of the criterion measure rather than the "meaning" of test scores (Bass & Turner, 1973, p.109).

Findings from differential constructs studies (e.g. Campbell, 1973; Cascio & Valenzi, 1978; Kraiger & Ford, 1990) indicate that correlations between ratings and various indices of performance are sometimes moderated by race of the ratee, race of the rater, or both. This research suggests that, relative to black ratees, there are more (frequency) intercorrelations among dimensional ratings and possibly halo for white ratees; ratings of overall performance are more correlated with dimensional ratings; and

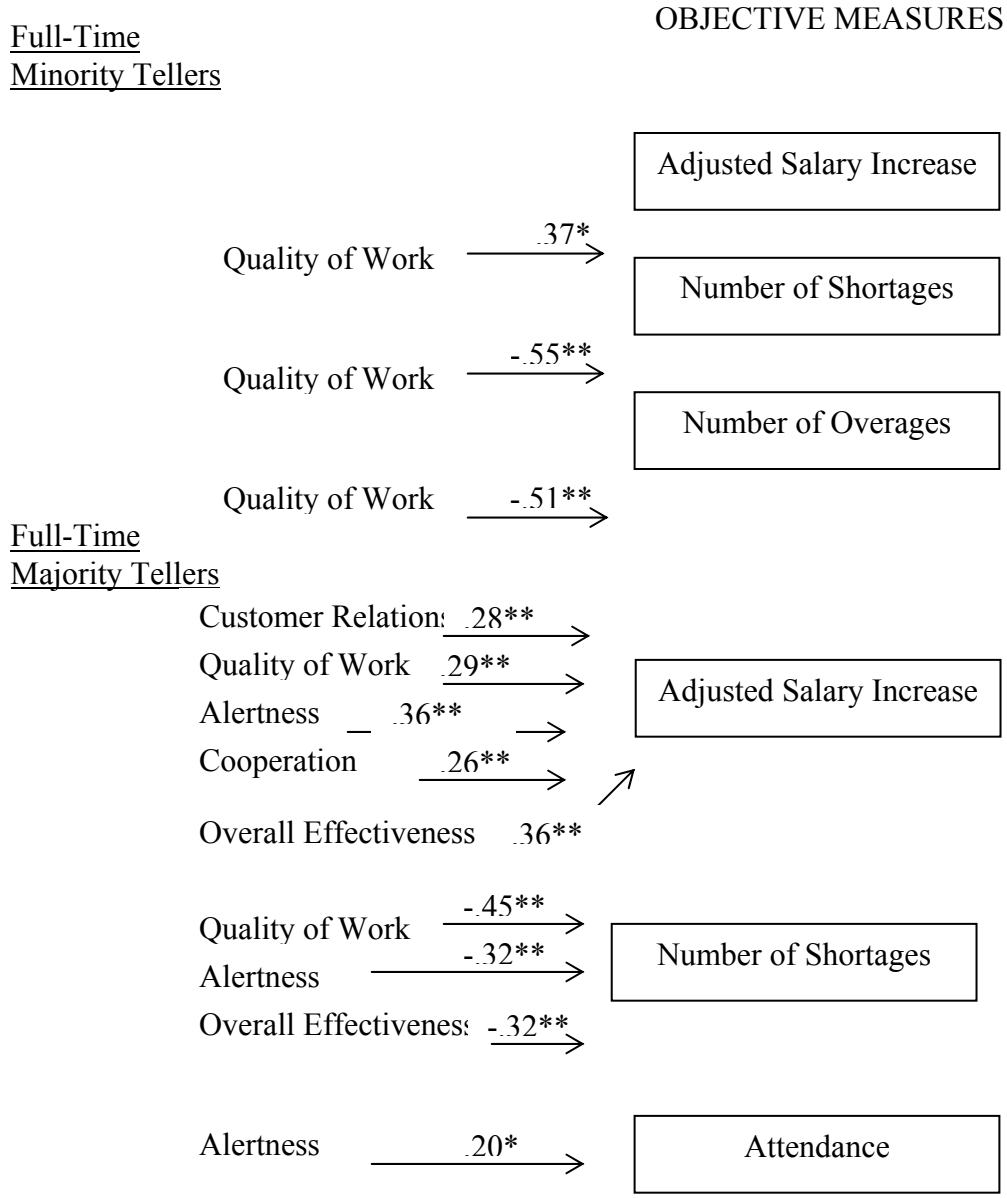

Figure 1. Correlates of objective performance indices for full-time minority and majority tellers (Bass & Turner, 1973).

Note. *p<.05. **p<.01.

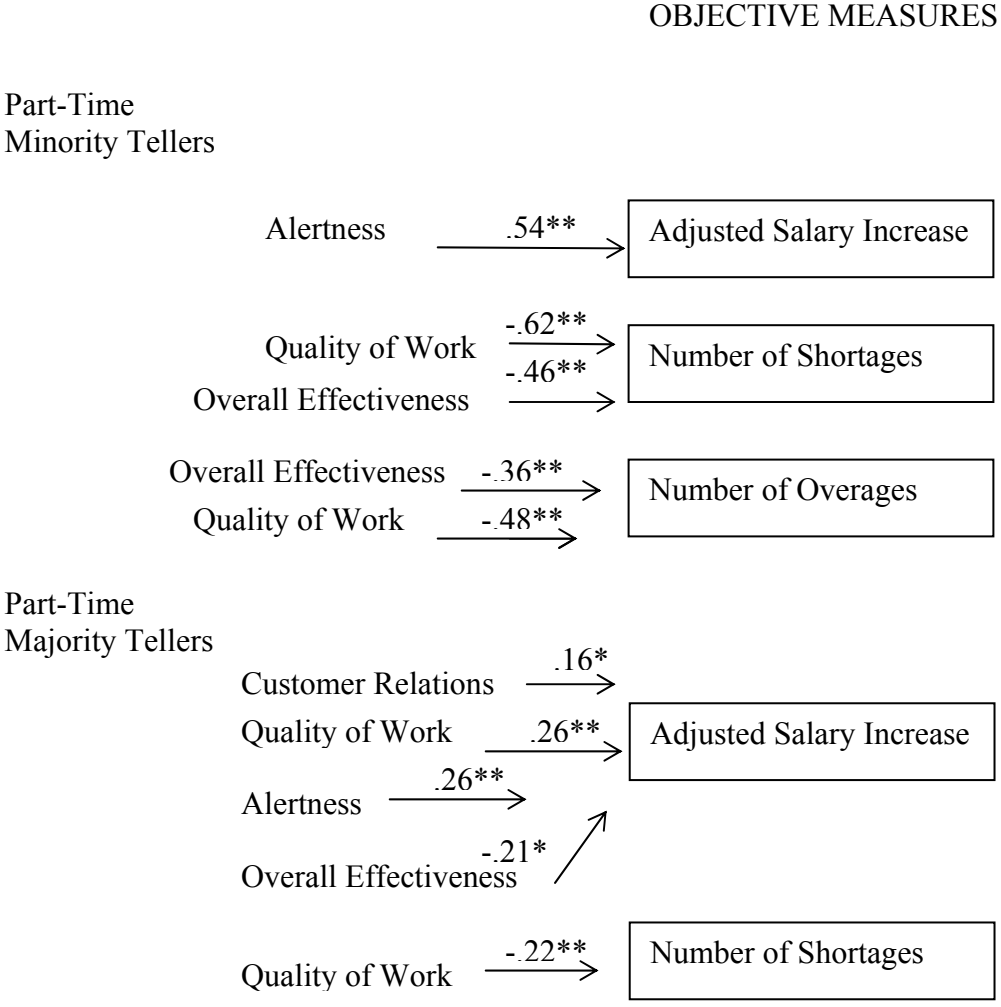

Figure 2. Correlates of objective performance indices for part-time minority and majority tellers (Bass & Turner, 1973).

Note. *p<.05. **p<.01.

objective and subjective criterion measures show more correlations for white ratees, while black ratees appear to be evaluated on a more limited set of factors. It is difficult, however, to draw definitive conclusions due to the inconsistency of results, and the incompleteness of sets of variables used as correlates in these studies (Oppler et al., 1992). In reviewing this literature, comparisons of findings were difficult due to the many different definitions of the criterion and quite often the lack of specificity in its definition. Kraiger and Ford's (1990) meta-analysis (described previously) of 14 studies that employed a differential constructs approach was a useful contribution. They found that ratings of minorities (African Americans) were more strongly related to objective measures of performance than ratings of the majority group; leading them to repeat their call (Kraiger & Ford, 1985) for research that examines the contextual and process variables that underlie group differences in ratings (Kraiger & Ford, 1990).

Models of Performance Ratings

Another approach to understanding influences on performance ratings has involved the use of statistical methods such as path analysis to identify the relative impact of different factors in arriving at performance judgments. Although these models were not developed to examine group differences in criterion performance, they have been applied to the study of predictor differences across groups. To this extent, they may provide some understanding of what predicts performance ratings for different groups. The most notable early studies are those conducted by Hunter and his colleagues (Hunter, 1983; F. L. Schmidt, Hunter, & Outerbridge, 1986). Using a military sample, Hunter (1983) developed a causal model of performance that includes cognitive ability, job knowledge, job proficiency and supervisory performance ratings. He found that the

effect of cognitive ability on ratings is indirect; that job knowledge has a direct effect on ratings; and job knowledge and task proficiency mediate the effects of cognitive ability on ratings. According to Schmidt and Hunter (1998) general mental ability is a good predictor of job performance because more intelligent people acquire job knowledge more rapidly, acquire more of it and this causes their performance to be higher. Schmidt, Hunter and Outerbridge (1986) extended the model to include job experience as a determinant of job knowledge and job proficiency. Job knowledge mediated the relationship between cognitive ability and supervisory rating.

Later studies have expanded this model to include motivational aspects of performance. Like the earlier models, these studies did not involved examination of group differences. Borman, White, Pulakos and Oppler (1991) used a sample of first-term soldiers (Project A data) to both test and expand Hunter's (1983) model. In addition to Hunter's explanatory variables which can be considered maximal "can do" cognitive measures (Borman, White, Pulakos, & Oppler, 1991), they added other variables that reflect "will-do" motivational measures: achievement orientation and dependability; and disciplinary actions, awards or commendations received. They also tested the stability of the models across jobs. The results of their replication are shown in Figure 3. Hunter's model was partially confirmed in that ability impacted ratings through job knowledge and task proficiency. There was no direct path between ability and ratings. However, unlike Hunter's findings, there was no direct relationship between ability and task proficiency. These findings suggest a completely mediational model. The variance in ratings accounted for was .14 after correction for attenuation (compared to .16 for Hunter's model). Borman, White, Pulakos and Oppler's (1991) revised and extended model

accounted for more than twice the variance in ratings than Hunter's original model (R^2=.31). A more parsimonious model of the cognitive measures was tested which resulted in a better fit. This was used as the starting point for the expanded model which is shown in Figure 4. Task proficiency and disciplinary actions had large direct effects. The ability→job knowledge path showed a strong indirect effect through task proficiency which highlights the role of cognitive ability in task performance (Borman et al., 1991). Dependability had a large indirect effect through disciplinary actions. Similarly, Borman, White and Dorsey (1995) found a completely mediational model where inclusion of interpersonal factors in a performance rating model increased the variance accounted for from 13 to 28 percent. These findings show that temperament factors can affect maximal performance measures or contextual factors can affect task factors.

Lance and Bennett (2000) had similar findings. Consistent with Borman et al. (1991, 1995), they found a complete mediational model with job knowledge and job proficiency mediating the relationship between cognitive ability and supervisor rating. Task experience and job experience also had direct effects on job knowledge and job proficiency. However, only task experience had a direct effect on rating. Job experience had no direct effect. Also consistent with Borman et al. (1991), disciplinary actions and awards had direct effects on rating. Thus task and motivational factors affected supervisor ratings.

Research has examined the fit of these models for different groups. In an extension of Borman et al.'s research, Pulakos, Schmitt and Chan (1996) evaluated the fit of a similar model of supervisory performance ratings for gender and different ethnic groups. Their findings suggest that raters may weigh performance factors differently for

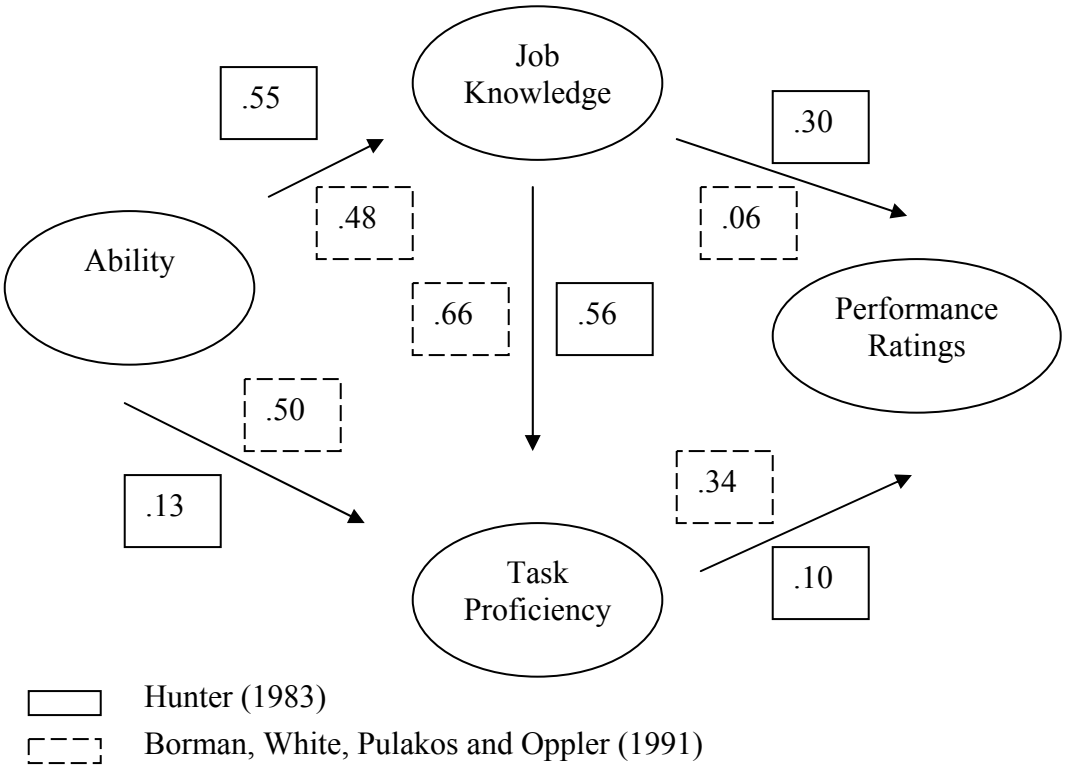

Figure 3. Hunter's (1983) model of performance ratings with findings from a replication (Borman et al., 1991)

different groups. Variables included in this model were cognitive ability, job knowledge, written job sample, role play job sample, motivation, practical intelligence (a dimension shown to result in smaller group differences than traditional general ability measures (Pulakos, Schmitt, & Chan, 1996)) and supervisor ratings. They found no differences in model fit for gender and racial group although the adjusted goodness of fit was lowest for African Americans (.83). It was "well above .90" for all others (p.115). Sample size for the African American subgroup in particular was a limitation in terms of detecting significant differences in the overall test of differences in subgroup models (Pulakos et

al., 1996). Also, there were differences in parameter estimates for the different group analyses suggesting that raters may weight factors differently depending on the subgroup.

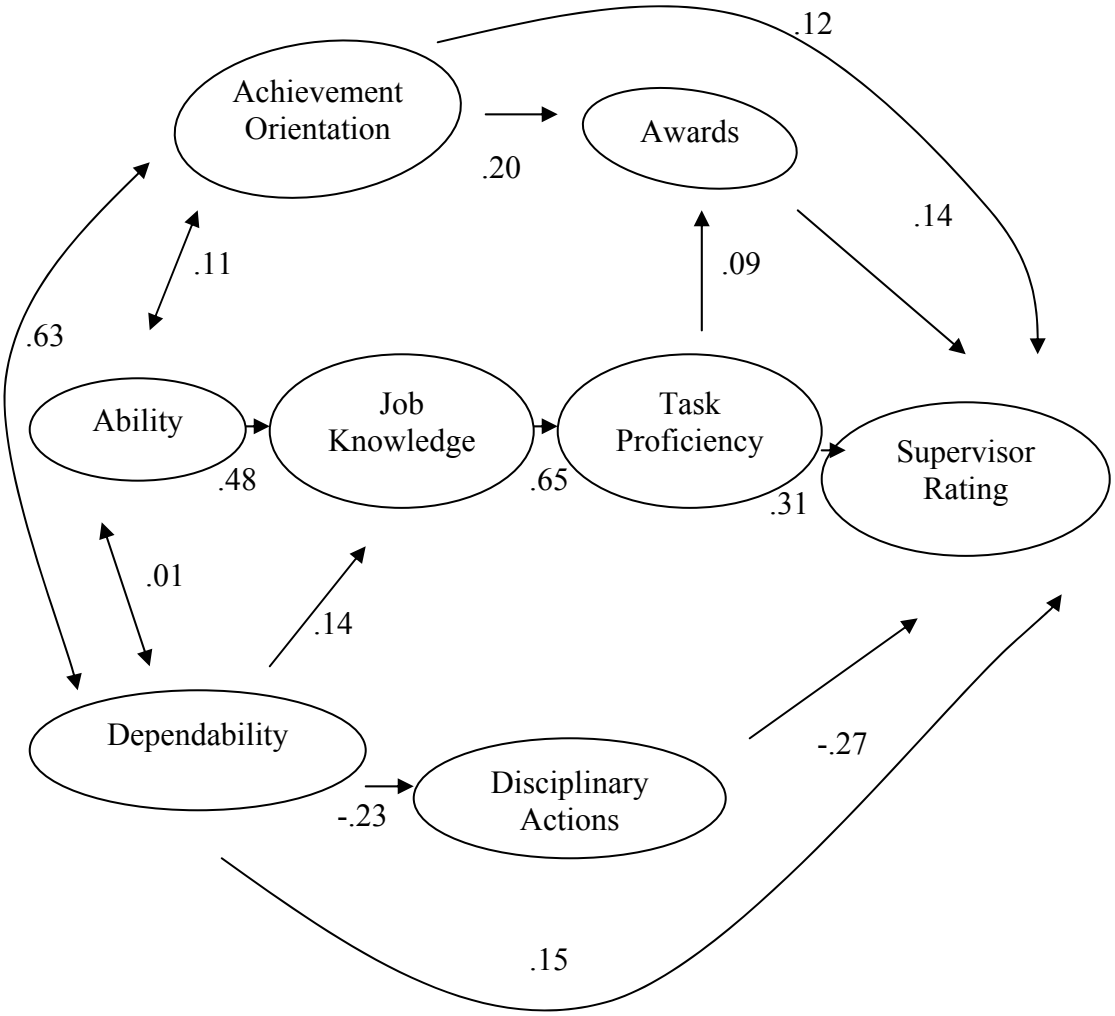

Figure 4. An expanded model of performance ratings (Borman et al., 1991)

Specifically, cognitive ability had the largest impact on the written job sample for African Americans (.40). This was the largest effect for this group, whereas the largest estimate for the white sample was the cognitive ability to job knowledge path (.48).

Pulakos, Schmitt and Chan's (1996) findings differed from the earlier study (Borman et al., 1991) in that job knowledge had a direct effect on ratings although the impact is substantially smaller than the ability→job proficiency link. Performance on written job samples and motivation had less impact on performance ratings. R^2 (.11) was comparable to the earlier studies (Borman et al., 1991; Hunter, 1983). These findings are also consistent with Campbell's (J. P. Campbell, Gasser, & Oswald, 1996; J. P. Campbell et al., 1993) general model of performance.

These studies confirm that motivational and temperament factors affect supervisor ratings and that ratings do not occur in a sterile context as the earlier models by Hunter and his colleagues would suggest.

Conspicuous by it absence in these studies is a discussion of the criterion used. Emphasis has been placed on identifying predictors to the exclusion of any examination or detailed description of the ratings used. Ratings are taken as sacrosanct even though one of the key researchers recently notes that in some settings where supervisors have limited opportunity to observe subordinates, ratings are potentially less accurate (Schmidt, 2002).

Wherry's theory of rating states that there are three types of factors that influence performance ratings: the ratee's actual job performance, rater biases in the perception and recall of that performance, and measurement error. In a test of this theory, Lance (1994) concluded that "ratings were stronger reflections of raters' overall biases than true performance factors" (p.768). He found that raters' idiosyncratic tendencies (all effects associated with individual raters including halo and leniency) accounted for more variance than actual ratee performance. His results were replicated by Scullen, Mount

and Goff (2000) who found that idiosyncratic rater effects accounted for 43 and 51 percent of the variance in "boss" ratings in the two data sets used. Corresponding figures were 64 and 52 percent for peer ratings. Actual ratee performance (general and dimensional variance), in comparison, accounted for only 30 percent (average of 28 percent for peers) of the total variance in both data sets. Thus, a greater proportion of variance in ratings is associated with biases of the rater than with performance of the ratee (Scullen, Mount, & Goff, 2000). Scullen et al. (2000) point out that although their study does not investigate the nature of individual/idiosyncratic effects, research of this type is needed including research that investigates influences such as racial and gender biases.

Significant race effects have been identified in job knowledge, task proficiency, and supervisor ratings. However, there are inconsistencies in race effects across different criteria and predictors vis â vis supervisory ratings. For example, there are larger race effects for objective measures of job knowledge (consistent with race differences in cognitive ability measures) than in subjective measures of the same criterion (Ford et al., 1986). Conversely, there are larger race effects in subjective measures of task proficiency than in objective performance measures (Ford et al., 1986).

It is not clear what accounts for the smaller race effects in subjective ratings of job knowledge relative to the larger effect for objective performance on the same criterion or what accounts for the lower supervisor ratings on task proficiency relative to actual performance on the same criterion. Similarly, if hands-on task proficiency is directly related to supervisory ratings (Borman, White, Pulakos & Oppler, 1991; Lance et al., 2000), and there are small race differences on actual (objective) performance (Ford et

al., 1986), then what accounts for the lower supervisory ratings? Finally, if there are large differences in job knowledge and pre-hire aptitude tests, what accounts for the relatively smaller mean differences (Sackett & DuBois, 1991) in supervisor ratings?

Summary

Current approaches to the study of subgroup biases in evaluation of criterion performance have been able to identify rater and ratee race effects, but these approaches are limited in their ability to distinguish criterion contamination from true performance differences. The inconsistencies in race effects across different criteria compared to supervisory ratings on the same criteria reinforce the need to understand the underlying meaning of performance ratings and any potential differences across groups. It has been argued that examination of subgroup differences on various criteria and studies of relationships among various criteria will not provide answers to the question of bias since we do not have the ultimate criterion (Schmitt & Noe, 1986). However, a differential constructs approach, which involves examination of the construct validity of ratings according to group membership of the rater or ratee, can provide insight to whether these criteria are being differentially applied by raters across different ratee groups. Before outlining the specific objectives of the present study, the next section addresses the theoretical explanations for group differences in job performance. This provides the conceptual framework for this study.

Theoretical Explanations for Race Differences in Performance Evaluations

There is very little theory regarding ethnic group differences in job performance. The few explanations that have been posited are generally untested, and the existing theoretical explanations are borrowed from gender research.

The primary exception to this is Dipboye's (1985) "holistic" model. Dipboye (1985) has proposed a holistic model in an attempt to explain differences in appraisals of different demographic groups (men and women, young and old, blacks and whites). He has criticized research on group differences in appraisals for an "overdependence on the stereotype-fit model and passive-observer research methods" (p.117).

The stereotype-fit model holds that raters attribute requirements to a position that are consistent with their stereotype of successful incumbents. A ratee's performance is favorable to the extent that the rater's perceptions of the individual fit his/her stereotype of the job. This operates through a filtering process whereby information consistent with a rater's expectations for a given ratee is more likely to be processed than information that is inconsistent with these expectations. Raters then compare the behavior of "the ratee to the stereotype of the ideal incumbent to form an opinion of the ratee's fit to the job" (p. 117). This is consistent with research on the antecedent conditions of stereotyping which includes the perceived lack of fit between the ratee's category and occupation (Fiske, Bersoff, Borgida, Deaux, & al, 1991; Heilman, 1983). Passive observer research typically presents a hypothetical ratee and performance is assessed holding constant or varying objective levels of performance (Dipboye, 1985). Unfair bias

is operationalized as different ratings across groups at the same level of performance. There are obvious limitations to the external and internal validity of such paradigms. These include findings which are not generalizable to experienced raters; and the continued use of ANOVA in making group comparisons with no focus on individual differences in raters such as attitudes and stereotypes (Dipboye, 1985).

Dipboye argues that although stereotypes are important, the social, behavioral and affective determinants of unfair discrimination have been neglected. He proposes a holistic model that incorporates these elements within the stereotype fit model. Within this model, bias in evaluations occurs in the form of disliking for the ratee (affective), self-fulfilling prophecies (behavioral), and conformity to social pressures (social). This framework goes beyond simple group comparisons and explores a limited set of affective, cognitive and behavioral factors (Dipboye, 1985).

Dipboye's model is largely untested although there are some data that provide support for some of its propositions. Stereotype-fit suggests that compared to the majority group, performance ratings given to minorities will have greater relationships with factors negatively related to performance such as absenteeism and accidents (Dipboye, 1985). Since expectations of ratees are associated with salient characteristics such as race of the ratee, the types of information that influence ratings given to different racial groups should vary (Dipboye, 1985; Oppler et al., 1992). Studies on the consequences of stereotyping show that negative attributes of stereotype category members are exaggerated while positive attributes may be discounted (Fiske et al, 1991).

One empirical test of stereotype-fit, described earlier, failed to find a relationship between supervisory performance ratings and negative performance indices for minority

ratees but found partial support in peer ratings (Oppler et al., 1992). Conversely, Bass and Turner (1973) found that majority rater ratings of performance were more highly correlated with performance errors (negative performance indices) such as 'number of shortages', 'number of overages' and 'attendance' for minority than majority bank tellers. Stereotype-fit might partly explain the findings of Ford et al. (1986) described previously (Oppler et al., 1992). They found significant race effects in ratings of performance indices where there were small differences in objective performance on the same criterion. As operationalized, the performance indicators included both positive (units produced) and negative (accidents, customer complaints) indices of performance. The race effects in ratings may have been due to higher correlations with the negative performance indicators for minorities compared to majority ratees. According to stereotype-fit, ratings for minorities should be more highly correlated with the negative performance indices relative to ratings for non-minorities.

In a more basic study, Tomkiewicz, Brenner, & Adeyemi-Bello (1998) asked a sample of managers to describe whites in general, African Americans in general, and successful middle managers using Schein's (1973; 1975) descriptive index. They found a large significant correlation between the profile (ratings) of whites and successful managers ($r=.54$, $p<.01$) and a non-significant correlation between the profile of African Americans and successful managers ($r=.17$). They conclude that managers are perceived to possess characteristics more commonly associated with whites than with African Americans.

The affective component of Dipboye's theory is consistent with more recent work on the impact of liking on performance ratings (Cardy & Dobbins, 1986; Cardy &

Dobbins, 1994; Lefkowitz, 2000; Lefkowitz & Battista, 1995). In a literature review of studies that examined the relationship between supervisory liking for subordinates and performance ratings, Lefkowitz (2000) concludes that supervisors' affective regard for a subordinate is frequently associated with "higher ratings, a higher quality relationship, less inclination to punish poor performance, and greater halo and less accuracy" (p.69). However, the direction of the causal relationship and the extent to which liking represents bias is still unclear (Lefkowitz, 2000). This lack of clarity is due to conceptual and methodological problems in the current research including the multiple and inadequate definitions of liking as well as a failure to recognize the developmental nature of supervisor--subordinate relationships. He proposed a causal model of affect (liking) and performance appraisal ratings wherein supervisory affective regard does not necessarily represent bias. However, he concedes that if ratings are impacted by social judgments such as liking, which in turn prove to be race or sex related, the result would be unfair bias. Although not directly related to subgroup differences in performance, Lefkowitz's model takes account of non-job relevant personal attributes of the ratee such as ethnicity and gender which can affect liking and consequently impact performance ratings. This effect could occur through the use of affect-based schemata, which structure the cognitive processing of performance information (Lefkowitz, 2000), and positive performance attributions (Cardy & Dobbins, 1994). Per Lefkowitz, four classes of variables may directly impact liking which has a direct impact on supervisory ratings: Similarity of the supervisor and subordinate (demographic similarity, opinions, attitudes, values, economic status, degree of mutual liking and trust); non job-relevant personal attributes of the subordinate (ethnicity, sex, political views, attractiveness); extra role behaviors

(organizational citizenship, impression management); and quality of the dyadic relationship. Of these factors, two (non-relevant personal attributes and supervisor/subordinate similarity) represent an indirect source of bias through supervisor liking (Lefkowitz, 2000). Extra-role behaviors and quality of relationship may be relevant or represent bias depending on the particular indicator or circumstances (Lefkowitz, 2000). Liking may also be based on job performance and other work-related behavior although this causal effect has never been tested directly. There is empirical support for some causal paths (e.g. rater-ratee dissimilarity and performance judgments, Haertel et al., 1999) although Lefkowitz's model is primarily untested.

Hunter and his colleagues provide an explanatory framework for ethnic groups differences in performance ratings, which is grounded in validation research rather than criterion theory. Based on their empirical model of performance ratings, they contend that criterion performance differences are a result of differences in cognitive ability across groups (Schmidt & Hunter, 1998). They further argue that these differences reflect true differences across groups and are not the result of rater bias. This conclusion is drawn primarily from research conducted with minority groups in the United States. Hunter's work has been criticized, however, for failing to reflect the multifaceted nature of the evaluation process and completely ignoring the social, situational, affective and cognitive elements of the process (Ferris, Judge, Rowland, & Fitzgibbons, 1994).

In Schmidt and Hunter's model, experience and ability predict supervisor ratings. It is argued here that the primary problem with this model is that although experience and ability may be valid predictors of performance, ratings do not necessarily represent valid measures of performance. The plethora of research that focuses on reducing rating errors

and enhancing the accuracy of performance ratings is testament to the fact that the rating process is far from perfect (see Borman, 1991 for a summary of issues).

With the exception of Dipboye's work, the existing theoretical explanations for group differences in appraisal ratings are borrowed from gender research. One such example is the treatment of the topic by Ilgen and Youtz (1990). They summarize potential sources of race effects on performance evaluation drawing from gender research in their illustrations. They propose two possible explanations for observed differences in criterion performance. The first is systematic rater biases that serve to elevate or depress performance ratings due to subgroup membership (Ilgen & Youtz, 1990) and the second is performance differences that stem from differences in ability based on the sorts of experiences minorities may have at work.

In terms of the first possible explanation, systematic rater biases, these include biases such as attributions, stereotypes, information selection and use, and judgment processes and stimulus saliency. These areas represent significant bodies of research on the impact of these processes on appraisals in general. Empirical treatments have primarily used majority samples or reflect gender research. Attribution research, for example, has focused on the role of attributions in organizations and their impact on behavior. One review (O'Leary & Hansen, 1983) cites a series of studies in which men's and women's successes on a task were perceived as having different bases. A man's success on a task is generally attributed to skill, whereas a woman's success is attributed to luck or effort (Deaux, 1976; Deaux & Emswiller, 1974; O'Leary & Hansen, 1983). Such findings have been hypothesized to generalize to minority group members (e.g. Ilgen & Youtz, 1990) but empirical data are lacking.

Similarly, research on stereotypes in performance evaluations have been studied from a gender perspective. The considerable research in this area suggests that stereotypes and sex role stereotypes, in particular, can play a significant role in the evaluation and perception of women at work (e.g. Fiske et al, 1991). The American Psychological Association (APA) filed a Brief Amicus Curiae in the U.S. Supreme Court in 1989 in the case of Price Waterhouse vs. Hopkins supporting the validity of research on stereotyping and its effects on evaluations at work (American Psychological Association, 1991). This was the first use of psychological evidence about sex stereotyping by the Supreme Court. Although the effects of sex-role stereotyping on evaluations of women have been well documented, there has been less research focus on raced-based stereotyping in evaluations. This lack of research was noted recently (see Boyce, Pratt, Bauer, Amelio, & Baltes, 2002).

The 1980s saw an emergence of theories of cognitive processes in organizations. Feldman (1981) presented a cognitive model of the performance appraisal process in which he argued that observers (raters) engage in a categorization process in making appraisal ratings. (DeNisi, Cafferty, & Meglino, 1984) similarly presented a model of the appraisal process based on the social-cognitive literature. Their model describes the method by which raters collect, encode, store, and retrieve information from memory, and the method by which he or she weights and combines this information in forming a judgment and, ultimately, arriving at an evaluation. These models served as a springboard for a number of research propositions regarding the role of cognitions within the performance evaluation process. Yet, it was not an area directly applied to minority group differences in performance evaluations.

Ilgen and Youtz's second explanation for observed differences in criterion performance is that differences in performance level stem from differences in ability or the sorts of experiences minorities might have at work. Ilgen and Youtz (1990) describe what they call "the lost opportunities effect" (p.271). Essentially minorities may have fewer and less favorable opportunities at work compared to their majority counterparts, resulting in lower performance. These lost opportunities may take the form of a lack of sponsorship or absence of role models in the organization; or result from factors such as composition of the workgroup, ingroup/outgroup relationships, and the extent to which minorities are seen as tokens. There is some empirical support for this argument. In one study, a portion of the race effect on job performance evaluations was shown to operate indirectly through two other variables – level of job discretion and organizational acceptance. Thus (per self reports) black managers' lower performance ratings were partially attributable to accompanying lower levels of job discretion and organizational acceptance (Greenhaus et al., 1990). African American subordinates with white supervisors have been shown to experience less supervisory support, developmental opportunities, and procedural justice than those with African American supervisors (Jeanquart-Barone, 1996).

Differential treatment and experiences of minorities in organizations undoubtedly impact development and possibly performance. This suggests that the "differential treatment" hypothesis (Ilgen & Youtz, 1990) is one tenable explanation for performance differences. Systematic rater bias, however, is an equally tenable explanation. Research findings such as rater-ratee race interactions in ratings (Kraiger & Ford, 1985); and differences across groups in the relationship between ratings and actual performance or

qualifications (e.g. Bigoness, 1976; Hamner et al., 1974; Powell & Butterfield, 2002; Schmitt & Lappin, 1980) suggests the role of some type of systematic bias. The rater bias and differential treatment explanations, however, are not mutually exclusive. They may both contribute to group differences and both deserve further study. The present research continues in the tradition of understanding the performance construct. The contribution lies with broadening the parameters of study in response to such calls (e.g. Dipboye, 1985).

In sum, there are potential theoretical frameworks for understanding group differences in criterion performance. However, there is no coherent theoretical explanation. The existing literature more accurately represents streams of empirical research seeking to explain these differences rather than a unified theory or framework.

Two elements of Dipboye's (1985) holistic model, stereotypes and affect, are useful in exploring potential differences across groups in the underlying meaning of performance ratings. Although Dipboye criticizes existing research for an over-dependence on stereotype-fit, very little research has been conducted on race-based stereotypes and performance evaluations. The focus of existing studies has been on gender-based stereotypes in appraisals (e.g. Dobbins & Russell, 1986; Robbins & DeNisi, 1993). It is argued here that stereotyping is pervasive. Indeed, fourteen years after the APA filed an Amicus Curiae Brief in the Hopkins vs. Price Waterhouse case, the APA has filed another Brief Amicus Curiae in support of the University of Michigan's admissions process (American Psychological Association, 2003). In this Brief they argue that research on associative processes "demonstrates conclusively" that unconscious stereotyping is widespread; and that although unconscious, these processes have

"significant real-world effects" including discriminatory behavior and judgments (Brief, 2003; p. 9).

Stereotyping is a by-product of normal cognitive processes, in particular, categorization (Fiske, 1998). Categorization is the process of "ordering the environment in terms of categories… through grouping (people), objects, and events as being similar…in their relevance to an individual's actions, intentions, or attitudes" (Tajfel & Forgas, 2000; p. 49). They provide cognitive shortcuts and serve the need to reduce the complexity of the social environment (Tajfel & Forgas, 2000). Categorization guides our search for new information, directs our attention to specific behaviors, and affects our memory for events and our distribution of rewards (O'Leary & Hansen, 1983) – each with significant implications for appraising performance. Stereotyping represents one form of social categorization. A stereotype is "a well-learned set of associations that link a set of characteristics with a group label" (Devine & Elliot, 2000). Stereotyping involves the use of category-based processing of information about the target individual. Once categorized, category membership is used to make inferences about the target person (Parsons, Liden & Bauer, 2001). Like categorization, stereotypes "simplify the social environment, expedite judgments, and free up cognitive capacity for other ongoing tasks" (Operario & Fiske, 2001) often at the expense of accuracy and fairness (Fiske, 1993).

It is posited here that raters possess "idiosyncratic theories" regarding effective performance (Klimoski & Donahue, 2001, p.25). This is consistent with research on implicit or folk theories of performance (Borman, 1987) which suggests that raters form categories of effective and ineffective performers which they use as a cognitive shortcut

in making performance judgments. These categories of effective performers can be influenced by stereotypes through social categorization (Barnes-Farrell, 2001) processes.

It is further posited that stereotypes and supervisory affect towards the ratee will impact supervisory ratings through their influence on supervisors' implicit (or idiosyncratic) theories regarding effective performance. Their impact will be reflected in the use and emphasis of different factors in evaluating minority versus majority performance. The following section outlines the objectives of this study.

Research Aims

The present research will examine underlying dimensions of performance ratings across groups. We know that based on experience and observation, raters possess implicit theories of performance and use these theories to evaluate performance (Borman, 1987). Further, these implicit theories of performance can impact the relative contribution of different factors to overall evaluation (Johnson, 2001). It is suggested here that stereotypes and supervisory affect will impact ratings through their influence on raters' theories of performance. Different theories of performance will result in the emphasis of a different set of factors across groups in arriving at an overall evaluation. This would provide evidence of differences in the underlying meaning of ratings across groups and evidence of possible criterion contamination in ratings.

A differential constructs framework will be used to explore the extent to which the underlying dimensions of ratings might vary by racial group. This broad research question will be addressed through group comparisons of: 1) the interrelationships among

performance ratings, and 2) the emphases given to different aspects of performance by supervisors in justifying their performance ratings of subordinates. Figure 5 illustrates the relationships to be examined.

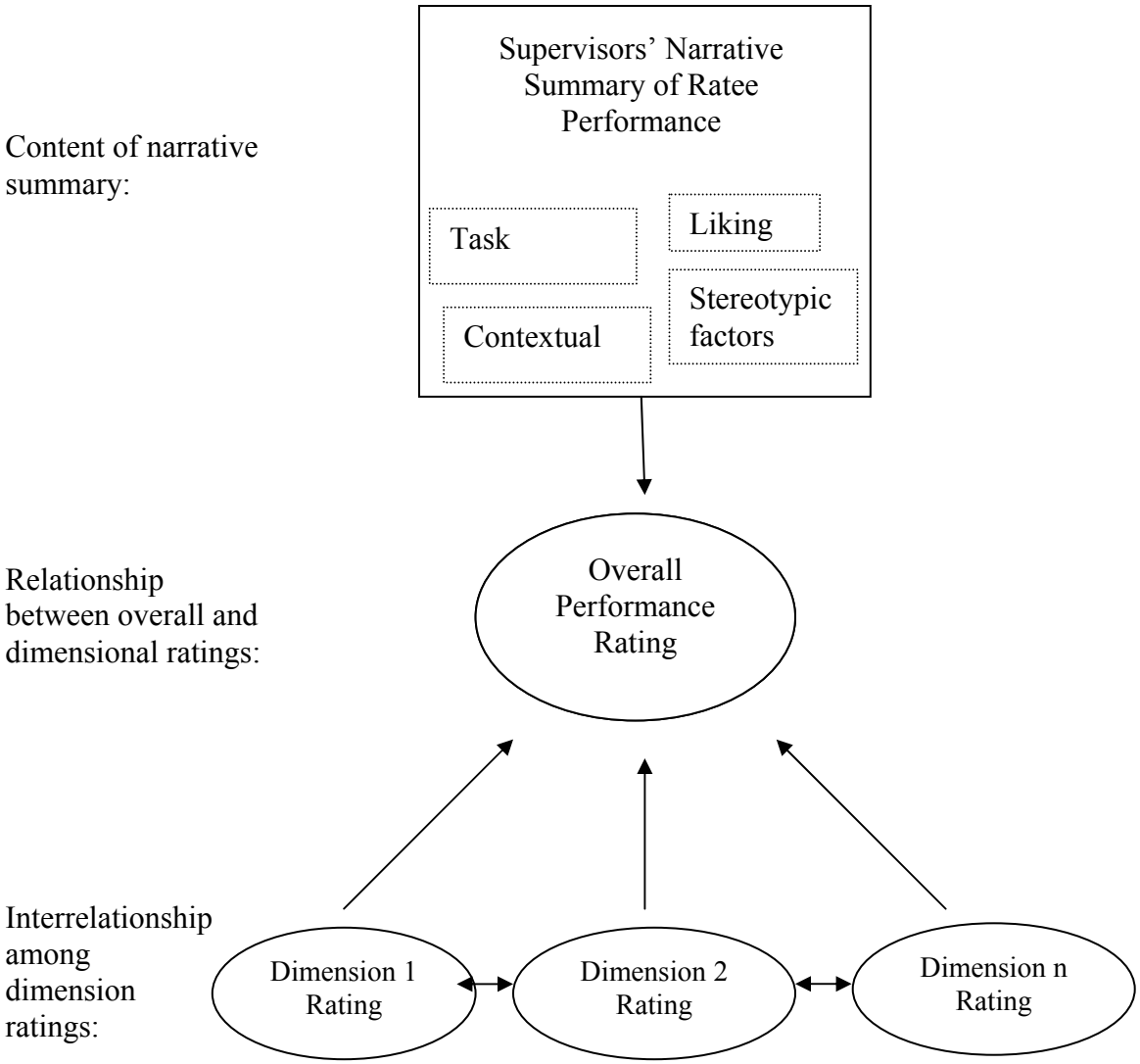

Figure 5. Relationships to be examined for each group.

Examination of the interrelationships among performance factors and their relationship to overall performance: A differential constructs approach would predict that performance ratings across groups have different correlates (correlations between ratings and other variables) and consequently different psychological meanings across groups (Oppler et al., 1992).

It is argued here that another operational test of the congruence in underlying meanings of performance ratings across groups would be the interrelationship among dimensional ratings of performance; and the relationship between dimensional and overall ratings of performance. The lack of congruence across groups in the relationship between dimensional and overall ratings may suggest that other factors (not included on the formal appraisal instrument) impact overall ratings. *Construct explication*, or "the process of making an abstract word explicit in terms of observable variables" (Nunnally, 1978, p.105), involves the examination of the relationship among measures. To the extent that ratings represent the same construct across groups, they should also share the same internal structure. That is, the internal consistency of ratings should be equivalent across groups. This reasoning is consistent with the basic principles of construct validation which state that "in the ultimate analysis, the 'measurement' and 'validation' of constructs can consist of nothing more than the determination of internal structures [of relevant measures] and cross structures [across those relevant measures]" (Nunnally, 1978, p.107). Measures of a construct should show evidence of homogeneity which would require that the items be generally intercorrelated (Cronbach & Meehl, 1955). Furthermore, performance dimensions are typically correlated. The use of uncorrelated dimension scores in policy capturing research in appraisal, for example, has been

criticized as threatening "the realism of the rating" as performance dimensions frequently have moderate to high correlations (Bass & Turner, 1973; Hobson & Gibson, 1983; Johnson, 2001, p.985).

At the most basic level, dimensional ratings should be correlated with overall ratings and should have a similar relationship with overall ratings within each group. Further, the relationships among dimensional ratings within groups should be similar across groups. We should not use a performance measure that represents different constructs across groups any more than we would use a predictor (e.g. a personality questionnaire) that represents difference constructs across groups.

From this framework, the underlying meaning of ratings will be examined through analysis of the relationships between overall and dimensional ratings, and interrelationships among dimensional ratings of performance for different ethnic groups.

Analysis of the content of supervisors' written summaries of subordinate performance and their relationship to rated performance: Supervisors' narrative summaries of subordinate performance (their justifications of ratings assigned) provide one indication of the meaning of performance for the supervisor. They may reflect what performance means to the supervisor, which may be different from the definition of performance provided by the organization. At the very least, they indicate the factors that were salient to the supervisor at the time of rating.

The research design for this analysis is shown in Figure 6. The nature and content of performance factors for majority and minority ratees will be examined on a number of facets guided by existing research on group differences in criterion performance and the multidimensionality of the performance construct. The focus will be on examining the

content of the performance factors (cells A, B, C and D) making comparisons across columns (good and poor performers), rows (minority vs. majority ratees), and cells within each group. Specific hypotheses are presented below.

	Good Performers		Poor Performers	
Minority Ratees	Performance Factors (A)		Performance Factors (B)	
	Positive Factors (E)	Negative Factors (F)	Positive Factors (G)	Negative Factors (H)
Majority Ratees	Performance Factors (C)		Performance Factors (D)	
	Positive Factors (I)	Negative Factors (J)	Positive Factors (K)	Negative Factors (L)

Figure 6. Research Design – Analysis of Written Summaries of Performance

Two theoretical explanations for group differences in performance ratings will also be tested here: stereotyping and affect (liking). Analyses will examine whether stereotypes and affect (liking) are reflected in supervisors' summaries of job performance.

Stereotypes: Stereotype-fit purports that appraisals reflect raters' perceptions of the fit between the ratee and the requirements or stereotypes of the job (Dipboye, 1985). Essentially, the rater compares his/her stereotype of the job requirements with his/her

perceptions of the ratee, which may include stereotypic views of the individual. When social stereotypes are invoked (based on salient characteristics such as race of the ratee), the types of information that influence ratings of different ethnic groups should vary based on raters' expectations of those groups. This model holds that minority ratings will have more relationships with negative performance factors. As discussed previously, this prediction is consistent with stereotype research which shows that negative characteristics of stereotype category members are exaggerated while positive characteristics are discounted (Fiske et al., 1991). Other stereotype research suggests that individuals may calibrate their standards based on negative stereotypes regarding minority groups' capabilities (Biernat & Manis, 1994; Biernat, Manis & Nelson, 1991), resulting in more positive evaluations in these low expectation domains (Harber, 1998).

Liking: Both theory (Lefkowitz, 2000) and empirical studies (Bernardin & Villanova, 1986; Cardy & Dobbins, 1994) suggest that supervisors' liking for subordinates directly impacts performance ratings even when objective performance level is partialled out (Harris & Sackett, 1988; Lefkowitz & Battista, 1995). The influence of liking cannot necessarily be construed as bias since the causal path between performance and liking has not been determined. However, the existence of liking as a performance factor is significant in itself given some job-irrelevant antecedents of liking, such as supervisor-subordinate similarity, gender, and ethnicity (Lefkowitz, 2000) that would constitute unfair bias or criterion contamination.

Using the design illustrated in Figure 6, analyses here will examine the content of the positive and negative factors reported by supervisors for minority and majority ratees

at different levels of performance (cells E to L) and their relation to the existing literature on stereotyping and liking in appraisal.

Exploration of the underlying meaning of performance ratings and the impact of stereotyping and liking on performance ratings will shed some light on whether ratings have different correlates for different groups and whether they may represent different constructs. This would constitute criterion contamination (Kraiger & Ford, 1990).

This study contributes to existing research in a number of important ways:

A key strength is the use of real performance data in a field setting. Much of the existing research in this area has been criticized for its lack of psychological fidelity to the real performance appraisal process (Bernardin and Beatty, 1984; Lefkowitz, 2000). In particular, the over-reliance on the "paper people" paradigm and the resultant lack of external validity (Guion, 1983).

The focus is on explanation of ratings and explication of the performance construct rather that identification of a race effect. Several researchers echo the need for research that moves beyond simply identifying systematic biases in ratings to research aimed at understanding the evaluation processes involved (e.g. Ford et al., 1986; Kraiger & Ford, 1985; Pulakos et al., 1989).

The focus is on issues directly relevant to the performance appraisal process and the concomitant implications for group differences. In his comments on gender research in performance appraisal, (Ilgen, 1983) has argued that the utility of sex-difference research in furthering our understanding of performance appraisal is generally limited although it does contribute to the sex-role literature in social psychology. He suggests that such research should start with issues more germane to the appraisal process and

move to the potential impact for sex differences rather than the other way around. Explication of the performance construct is a fundamental issue to I/O psychology and will serve to further our understanding of the appraisal process both at a general level and across groups.

In terms of the broader contribution to theory and practice:

It has been argued that criterion differences between majority and minority groups in the U.S. have been consistent with minorities' lower scores on predictors, particularly cognitive measures. This is particularly significant in terms of validity generalization which suggests equal validity of cognitive measures for all groups. However, there is little research that has actually looked behind these validity coefficients. An understanding of the underlying meaning of performance ratings is crucial to shedding light on this argument.

This research has significant implications for practice. It would have implications for personnel decisions, and all aspects of equal opportunity legislation.

Research Questions and Hypotheses

Research Question 1: Are there significant differences in performance ratings between minority and majority group members?

Hypothesis 1: Minorities will score lower on average on supervisors' overall ratings of performance compared to majority group members. Research shows that minorities are generally lower than majority group members on supervisory ratings of performance (Ford et al., 1986; Sackett & DuBois, 1991; Wilson, 1995). Operationally, mean overall ratings should be significantly different across groups.

Research Question 2: Is the relationship between dimensional ratings and overall ratings different across ethnic groups?

Hypothesis 2a: Groups will be rated the same on dimensional performance ratings while their scores on overall performance will be significantly different.

There may be larger group differences on global measures of performance than on dimensional ratings of performance. There is some evidence that race effects are more likely with the use of global evaluations of job applicants and not evident based on behaviorally-specific ratings (Brugnoli et al., 1979). Raters may rely on global impressions and categorization processes (Feldman, 1981; Lord & Maher, 1989) in making overall ratings; while ratings on specific performance dimensions may require more conscious or controlled information processing. There is some evidence, however, that categorization can occur at the dimensional level (Lord & Maher, 1989). Differences across groups in the relationships among dimensional and overall ratings would suggest different underlying meanings in ratings for each group. Specifically, no differences in dimensional ratings, in the presence of significant differences in overall ratings, would suggest different underlying processes on the part of the rater in the assignment of overall ratings across groups. This would suggest criterion contamination. Analyses would involve: group comparisons of mean differences on each performance dimension; group comparisons of mean overall dimensional ratings; and group comparisons of mean overall performance ratings.

Hypothesis 2b: There will be a different pattern of relationships between dimensional and overall performance ratings for majority than minority group

members. Findings from differential constructs studies show that ratings of overall performance are more correlated with dimensional ratings for majority than minority ratees (Cascio & Valenzi, 1978; Kraiger & Ford, 1990). Minority ratees appear to be rated on a more limited set of factors. Operationally, relatively fewer dimensional ratings are correlated with overall ratings compared to majority staff. This suggests different underlying meanings across groups. Correlations between dimensional and overall ratings will be compared for both groups.

Hypothesis 2c: There will be less dimensionality among skill ratings for majority than minority staff. Empirical studies show that there tend to be more intercorrelations among dimensional supervisory ratings for white than black employees (e.g. Bass & Turner, 1973; Casio & Valenzi, 1978). There are higher correlations for white ratees suggesting possible halo while black ratees appear to be rated on a more heterogeneous set of factors. Irrespective of any differences on overall performance, the pattern of intercorrelations among dimensions should be similar in both groups. Even in the case where lower dimensional ratings reflect true score differences rather than bias, intercorrelations for minority ratees would be expected to be equally high as those for majority ratees – means should simply be lower for the minority group. Lower intercorrelations among dimensions for one group relative to another suggests a different factor structure for each group. This implies bias. Intercorrelations among skill areas will be examined for strength and frequency. There will be more and higher intercorrelations among skill areas for majority than minority staff.

Research Question 3: Are there differences across groups in relative emphases of performance factors cited in justifying overall performance?

Hypothesis 3a: There will be fewer mean positive mentions of task factors; and more mean negative mentions of task factors in supervisors' summaries for minorities compared to majority group members. Preliminary research on ethnic group differences in task and contextual elements of performance suggest that minorities are rated lower on task performance while the findings for contextual performance are less clear (Hauenstein et al., 2002). If there are overall differences in the underlying meaning of performance across groups (supervisors' implicit theories of performance), this would indicate that the differences in task performance are more likely a representation of bias rather than true performance differences.

Hypothesis 3b: Supervisors will emphasize contextual elements of performance more for minority than majority staff in factors cited in justifying evaluation of performance. Supervisors have been shown to emphasize social factors more in ratings of minorities than majority staff (Beatty, 1973). These behaviors are reflected within the interpersonal citizenship category of contextual performance (Coleman & Borman, 2000). There will be proportionally more mention in minority summaries of interpersonal citizenship behaviors compared to majority ratees. Support for this hypothesis would be consistent with systematic bias through a focus on different aspects of performance as a function of group membership. The result would be a different underlying meaning across groups.

Research Question 4: Is there any empirical support for stereotyping as a theoretical explanation for differences in supervisory ratings of majority and minority performance?

Hypothesis 4a: There will be proportionally more negative performance factors mentioned in supervisors' summaries for minority than majority staff both in terms of positive and negative occurrences. Dipboye's (1985) stereotype-fit model predicts that minority ratings will have greater relationships with what he calls negative indices of performance such as attendance or 'overages' and 'shortages' for bank tellers (Bass & Turner, 1973); essentially indices that represent negative examples of performance rather than neutral or positive indices. For example, supervisors are more likely to mention 'attendance' in summarizing a minority ratee's performance. Attendance might be mentioned positively ("excellent attendance") or negatively ("attendance needs to improve"). Similarly, 'shortages' may be mentioned positively ("no shortages this quarter") or negatively ("$200 in shortages this quarter"). This would suggest the use of a different set of factors in evaluating minority performance relative to majority performance.

Hypothesis 4b: There will be significant differences across groups in supervisors' tendency to make positive comments overall in their written justification of performance. Operationally, there will be significant mean differences across groups in the total occurrence of positive comments.

Hypothesis 4c: Majority staff will have more 'liking' factors in their performance summaries than minority performers at the same level of performance.

Ethnicity has been shown to be significantly correlated with liking (Lefkowitz & Battista, 1995). Factors such as the degree of similarity of the

subordinate to the supervisor, in terms of demographic attributes (Ferris et al., 1994), values, and attitudes, influence the degree to which the subordinate is liked by the supervisor (Lefkowitz, 2000). Lefkowitz has emphasized the developmental aspects of supervisor-subordinate relationships and suggested that tenure be included as a categorization variable so that analyses are performed on cohorts homogeneous in ternure. Holding tenure and performance level constant, majority staff should have proportionally more positive liking comments that minority staff.

General Research Questions: In addition to these hypotheses, there is an additional question that would be useful in shedding light on any differences in supervisor evaluations across groups. Specifically, what are the underlying themes in supervisors' narrative comments in general and across groups?

CHAPTER 3

METHODS

Sample

The sample consisted of bank staff from a national bank in the United Kingdom (UK). This was an opportunity sample in that the bank was a consulting client. Appraisal forms were obtained for 667 individuals. No personal data were provided that would allow identification of individual employees. Appraisees were within the managers' own branch. There was a total of 101 raters with an average of 6 supervisees each. The number of supervisees ranged from 1 to a maximum of 18.

Jobs Sampled

Complete information on jobs is not available, however, the primary job title held by the sample was cashier (approximately 80 percent). There were also coin sorters in the sample. Both groups were combined due to the small number of coin sorters in each group relative to cashiers.

Overall performance ratings for three annual reviews were provided – the two annual reviews immediately preceding the written appraisals as well as the year of the written appraisals. These data were provided along with demographic information. Thus three years' worth of overall performance ratings were available. The data obtained are

summarized in Table 5 below. As different data were missing for some employees, missing data are shown in the data tables.

	Sample Size		
	Valid Cases	Missing	Total
Appraisal Forms provided (Year 3)	667		
Ratings Year 1	634	(269)	903
Ratings Year 2	658	(245)	903
Ratings Year 3	710	(193)	903

Table 5. Performance Data Obtained

In terms of demographic characteristics, there were eight ethnic classifications plus an "other" category (see Table 6). As these data were provided by a UK organization, the ethnic classifications reflect those in that geographic area. Due to the small sample sizes, the ethnic categories were collapsed along two lines: majority/minority and white/black/Asian. The minority category was comprised of all minority ethnic and racial groups except Chinese. Chinese staff were excluded from the analysis due to the small sample size. Analyses could not be performed separately for that group and they could not be sensibly clustered with the other Asian groups. The other Asians in the sample are from India, Pakistan and Bangladesh, which are all in the region of the Indian subcontinent.

	Percentage	N		Cluster Percentage
African	.9	(8)		
Caribbean	6.2	(56)	Black	8.4 (76)
Black Other	1.3	(12)		
Bangladeshi	.7	(6)		
Indian	4.8	(43)	Asian	6.1 (54)
Pakistani	.6	(5)		
Chinese	.6	(5)	Chinese	.6 (5)
White	70.3	(635)	White	70.3 (635)
Other	2.3	(21)	Other	2.3 (21)
Missing	12.4	(112)		

Table 6. Percentage of Entire Sample by Ethnic Group

To minimize the possibility of obscuring group differences given the number of different ethnic groups, staff were clustered along racial lines into Asian, black, and white staff. 'White' staff were those classified as white in the data base. The 'black' category was comprised of Africans, Caribbeans, and those who were categorized as 'black other'. The 'Asian' category was comprised of Bangladeshi, Indian, and Pakistani staff. The original classifications were provided by the bank.

It may be useful to put these data in context by providing some information regarding the percentage of ethnic minorities in the UK. In the UK, the term 'black and ethnic minority' includes individuals from African, Asian, and Caribbean ethnic backgrounds (Davidson, 1995). Minorities comprise five percent of the total population

in Britain (Jones, 1993). Within this five percent, the largest groups are Caribbean and Indian, approximately 25 and 26 percent, respectively (Jones, 1993).

Staff were predominantly female overall (66 percent) and moreso for the minority sample (79 percent). The average age of majority staff is 32. Minorities are slightly older with an average age of 34 (See Table 7).

	Total[a] Percentage	Majority	Minority
Female	66	63	79
Male	34	37	21
Average Age	32	32	34
N	765	635	130

Note. [a]Excludes missing values.

Table 7. Sample Characteristics – Gender and Age

As shown in Table 8, the average tenure for the majority group was 12 years and 13 years for minorities. Caribbeans and Indians had the highest mean tenure (14 years) and Africans and Chinese the lowest (8 and 3 years, respectively). Examination of means across the different classifications using a univariate analysis of variance (ANOVA) revealed significant differences in mean tenure across the ethnic groups. Correlations between tenure and overall performance rating were statistically significant for majority and black staff (see Table 9). These initial analyses resulted in the decision to statistically control this variable in later analyses.

Group	Mean (Years)	sd	F
Majority	12	8.3	2.004*
Minority	13		
African	14	7.9	
Caribbean	8	6.9	
Black Other	12	7.3	
Bangladeshi	10	8.4	
Indian	14	9.8	
Pakistani	12	8.2	
Chinese	3	3.8	
Other	9	8.9	

Note. *p<.05

Table 8. Tenure by Ethnic Classification

	Group		
	Asian	Black	White
Year 1 Ratings	.11	.41**	.16**
Year 2 Ratings	-.08	.26*	.22**
Year 3 Ratings	.17	.35**	.18**

Note. *p<.05 **p<.01

Table 9. Correlations Between Overall Ratings and Tenure by Ethnic Group

Performance Measures

The bank's annual appraisal form consists of two parts: a review of performance over the past twelve months; and what needs to be achieved over the next review period. The supervisor is provided with written instructions for completing the form. Part 1 consists of a written 'summary of performance'; an overall performance rating; ratings on specific skill areas; and a listing of areas of strength and development. In completing the written 'summary of performance', the reviewer is instructed (in the accompanying notes for completion) to summarize (in narrative form) his/her notes on the ratee's performance into a concise and full account of the ratee's achievements. The appraiser is asked to consider achievement in relation to the jobs' performance standards (for example, business development, quality standards, accuracy) and the broader organization standards (for example, open, honest). These broader standards are expected of all staff irrespective of the job they perform. They are instructed to include extra achievements and personal development plans. The summary is to describe performance and behavior rather than the individual. Appraisers are asked to avoid adjectives such as 'enthusiastic' and personality-related terms.

Overall performance is evaluated on a 5-point scale using the anchors: O=outstanding, H=high achievement, G=good performance, I=improvement required, and U=unacceptable. In the *skills* section of the form, appraisers are required to rate staff on 12 skill areas: business development, business awareness, written communication, verbal communication, customer service, decision making, initiative, numeracy, planning and organizing, team work and team leadership, use of systems, and work standards. The supervisor indicates the 'level required' and rates the 'level achieved.' Both ratings are

made on a 7-point scale ranging from 1 to 7, with 1 being the highest level and 7 being the lowest. These rating are done every 2 years rather than each annual review. Finally, supervisors list areas of strength and areas for development on designated areas of the form.

In Part 2 of the form, if performance is below the mid-point of the scale (good performance), the supervisor is required to complete a performance improvement plan. This is not required if the ratee performs above the mid-point of the scale. Otherwise, supervisors list specific development action to be taken over the next 12 months and a timetable for that action. In this section, ratees have an opportunity to comment on the appraisal. For confidentiality reasons, an actual form cannot be shown here. However, an outline of the appraisal form is shown as Appendix A.

Analyses

Performance Ratings

Means on overall and dimensional (skills) ratings were compared across groups with a series of analysis of variance (ANOVAs). Given the significant differences in tenure across ethnic groups and the correlation between tenure and overall performance rating for the majority and black samples, analysis of covariance (ANCOVA) was used in these comparisons to control for the effect of tenure on the dependent variables.

As described previously, skills were rated on a 7-point scale with 1 being high and 7 being low. Individuals were given a 'level achieved' rating relative to a 'level required'. The difference between the level of performance required and the level of

performance achieved was computed for each skill area. This approach is commonly used in studies for comparison of 'is' versus 'should be' ratings (R. C. MacCallum, personal communication, July 8, 2003). Thus the simple formula for this statistic was:

Difference = Skill Required – Skill Achieved

If an individual performed at a level higher than required, the difference was positive. If an individual performed at a lower level than required, the difference was negative.

Mean overall ratings were compared for each of the three appraisal years; and skills ratings were compared for year 3, the only year for which they were available.

One of the underlying assumptions of ANOVA is the homogeneity of variance across groups (Keppel, 1982). Violation of this assumption leads to an increased possibility of Type I error (Keppel, 1982). *Monte carlo studies* have shown that violations of homogeneity of assumptions have their smallest effects when sample sizes are equal (Collyer & Enns, 1986). This was a concern in this study given the unequal sample sizes. However, the problem can be addressed through the use of alternative F tests that take into account the differences in sample sizes and any heterogeneity of variance (Keppel, 1982). For each of the ANCOVA analyses, Levene's test of equality of error variances was computed to test equality of error variances across groups (Myers & Well, 2003). Variances and boxplots were examined for each group on each dependent variable. Generally, if the largest variance is only twice as large as the smallest variance, this is not considered a violation of the homogeneity assumption (Collyer & Enns, 1986). Also, bias due to suspected heterogeneity of variance can be minimized by replacing the .05 significance criterion with a more stringent test, a .01 criterion (Collyer & Enns,

1986). If it was suspected that this assumption had been violated, Welch's (F_w) test was used. Welch's test is a special F test that is recommended (Myers & Well, 2003) for use with unequal variances and takes unequal sample sizes into account in calculating F. It has been argued that unless populations are "seriously non-normal" (Kerlinger, 1973; p.287) or variances too disparate, it is preferable to use a parametric rather than a non-parametric test. It is not the intention to minimize the importance of homogeneous variance across groups in the use of ANOVA, however, it is worth noting it has been argued that the "importance of normality and homogeneity is overrated" (Kerlinger, 1973; p.287).

Narrative Data

There are two general approaches to analyzing text which occur at two different levels. One involves the analysis of words, while the other involves the analysis of blocks of text. Methods of analyzing words include techniques such as word counts, structural analysis, and cognitive maps. Approaches to the analysis of blocks of text involve coding at their core. The researcher must make sense of or derive meaning from the blocks of text. There are several theoretical approaches to this type of analysis including grounded theory, schema analysis, and classical content analysis (Ryan & Bernard, 2000). This study employed one method from each general approach: content analysis and word counts.

Content Analysis: Content analysis has been used to analyze text and can provide insights to written data that would be difficult to derive from other approaches (Erdener & Dunn, 1990). This technique has several advantages given the nature of the data being analyzed here and one goal of the present research which is to identify underlying themes

in supervisors' written summaries of performance. Among its advantages are (Erdener & Dunn, 1990):

1. It permits systematic interpretation of textual material based on objective criteria;
2. It can convert qualitative material to quantitative data … for further analysis … using statistical procedures;
3. It does not necessarily require large amounts of data, but can also be used for small-scale studies…;
4. It is the ultimate unobtrusive measure;
5. It can be used in combination with other research methods such an quantitative analysis of financial statements … to combine fine-grained with coarse-grained research methods in triangulating on complex issues or organization. (p.292)

The content analysis was conducted at the *manifest* level which captures the surface characteristics of the words, rather than the *latent* level which attempts to identify the underlying meaning of the text. The aim was to analyze the content in the language of the rater rather than identifying underlying meanings in the summary. Analysis of manifest content focuses on word frequency counts and occurrences of key words in relation to other words in the sentence (Erdener & Dunn, 1990). Measurement involves percent of total words, or ratios in comparison with the occurrence of other words in the text. The key limitations to this method are the likelihood of human error and the inherent subjective interpretation with its consequent impact on reliability and validity. These issues are being addressed through: interrater reliability analysis; the use of a more

inductive approach – reasoning from the parts to the whole; and linking the content to existing research on the criterion construct.

There were four basic steps to the content analysis: sampling, identifying themes, building codebooks, and marking texts (Ryan & Bernard, 2000). .

Sampling – 667 completed appraisal forms were obtained for analysis. A large sample was sought to allow the quantification of the qualitative material and comparison of themes of across groups. The section of the form identified for analysis was the "summary of performance" section completed by the supervisor. An example of this section is shown in Appendix B. The unit of analysis was words or word senses rather than sentences (Ryan & Bernard, 2000); essentially, words or two- to three-word short segments that reflect the supervisor's view of the individual and/or their performance.

Identifying themes – The goal here was to identify the factors that supervisors take into account in describing the performance of subordinates; in essence, the aspects of performance that they list in summarizing performance and documenting the overall rating given. The coders were looking for words that described actions of the subordinate, personal characteristics, and/or consequences of the subordinate's behavior. The process was inductive. No preconceived themes or concepts from the existing literature were used, the analysis involved recording the supervisor's descriptors.

Building codebooks – A coding dictionary was derived from a sample of appraisal forms. Two independent raters (one I/O psychologist and the present researcher) were given copies of the same 10 appraisal forms and asked to list the performance "factors" that supervisors used in describing each ratee's performance. They each produced a list of factors for each of the 10 appraisals. An example of a list of factors produced and the

corresponding 'summary of performance' is shown in Table 10. The raters then combined (independently) the factors across all 10 forms to produce one list of factors per coder.

Coding Summary – Form 1	
Name: *Sue Smith*	Job Title: *Cash Clerk*
Summary of Performance: *Achieved standards* *Significant improvement* *Productivity improvement (exceeds targets)* *Accurate* *Well done* *Cooperative* *Helpful* *Quick learner* *Stays late* *Helps out team* *Excellent assistant supervisor* *Commendations* *Sickness record needs improvement* *Initiative* *Can be left unsupervised*	

SUMMARY OF PERFORMANCE

Sue has successfully achieved all her key job standards and "ABC Bank" core standards. Sue has shown significant improvement during the second half of the year. Her productivity improved to target level and above on all transaction types. "Well done." Her difference ratio has also become very good. Sue is very cooperative and always helpful. She has become a quick learner and can cover all counting duties as a S2. Sue always stays late to help out the team effort. She is also an excellent assistant supervisor and has had two commendations on her evidence of performance. Her sickness record needs to improve. She can be left unsupervised and uses her initiative to good effect.

Table 10. Example of Coding Summary Form and corresponding Summary of Performance.

The two raters met to compare the lists they had generated independently. There was high agreement on the factors identified by both. There were only a few factors uniquely identified by each. After discussion, both sets of these factors were added to the list. This formed the dictionary for content analysis. This dictionary is shown in Table 11.

Able to work unsupervised	Focus	Pride
Accurate	Good rapport	Prioritization
Active in self developm't	Helpful	Productivity
Adaptable	Helps out with other tasks	Professional
Adjusted to change	High quality	Punctuality
Alert	High standard	Quick
Attendance	Honest	Quick learner
Attitude	Initiative	Referrals
Aware	Innovative	Relationship with cust.
Calm	Inquisitive	Reliable
Can-do	Integrity	Responds positively to challenges
Cheerful	Interpersonal skills	Sales
Commitment	Knowledgeable	Sickness record
Communications	Leadership	Social skills
Competent	Learns new tasks	Speed
Concentration	Makes suggestions	Stress tolerance
Confident	Mature	Supervision
Conscientious	Methodical	Takes on new tasks
Consistent	Motivated	Team player
Cooperative	Organized	Thorough
Cross Sell	Patient	Trains other member of staff
Diligent	Personal appearance	Versatile
Efficient	Personality	Very high standard
Enhanced results of unit	Polite	Works extra hours early or late
Enthusiastic	Popularity	Works hard
Experience	Positive	Customer service
Flexible	Potential for developm't	Awards/certificates
Friendly	Pressure	Praise (from manager)

Table 11. Dictionary for content analysis.

Marking text/summary sections -- Each factor was assigned a code. These codes served as tags (Ryan & Bernard, 2000) rather than values so represented nominal data at this stage. Examples of tags are shown Table 12 below.

Performance Factor	Code
Popularity/social skills/polite	10
Cooperative/helpful	11
Team player	12
Supervision/leadership skills	13
Communications	14
Friendly/cheerful	15
Helps out with other tasks	16
Good rapport	17
Calm, Patient	18
Relationship w/customer	19
Mature	20

Table 12. Examples of tags used to code performance factors.

If a factor was cited in a positive manner (for example, "John is a team player"), this was considered a positive mention of the factor and the tag assigned had a positive value. If the factor was cited in a negative way ("John is not a team player"), this was considered a negative mention of the factor and a negative sign was used in coding. Thus, the tag for the latter example would be '-12', whereas the tag for the former would

be ‘12’. Negatively and positively tagged factors were assigned different variable names. For example, p12 (team player) and n12 (team player negative). Frequency analyses were computed to determine the occurrence of each factor across the total sample. Factors with total frequencies (positive and negative occurrence of the factor) of 7 or below were dropped from the analyses due to their infrequent occurrence relative to the total sample size. Seven was chosen as the cutoff only because it appeared to be a natural cutoff. Frequencies tended to be below seven or substantially higher with only one or two factors having frequencies between 11, 14, 15, 18 and 19. Table 13 shows these data listed in descending order of frequency.

Supervisors overwhelming cited positive factors in justifying the performance ratings. The frequency of negative factors was so low that, for the most part, these were not used in subsequent analysis. Due to the sheer number of factors and the low frequencies for some, factors were subjectively sorted into clusters. Each factor was written on an index card and placed into a group based on similarity. The resulting clusters are show in Table 14.

Factors were also designated as task versus contextual elements of performance. These designations were made based on definitions of these dimensions in the literature. As discussed previously, task performance refers to task proficiency and reflects job knowledge, skills, abilities, experience and training (Hattrup et al, 1998). Contextual performance refers to behaviors that contribute to the maintenance or enhancement of the context of work. These are non-task behaviors (Organ, 1997). Those factors that did not seem to clearly fit into one or the other category were excluded from the later task/contextual analyses. Table 14 shows these designations.

Performance Factor	Positive Occurrences	Negative Occurrences	Total
Cross sell, Sales, Referrals	158	52	210
Accurate	163	43	206
Customer service	158	7	165
Speed	124	15	139
Cooperative/helpful	134	1	135
Helps out with other tasks	126	0	126
Popularity/social skills/polite	116	2	118
Friendly/cheerful	99	0	99
Knowledgeable/inquisitive	86	11	97
High standard/quality	85	4	89
Awards/certificates	81	0	81
Shows initiative	62	18	80
Works extra long hours early or late	77	1	78
Motivated/enthusiastic	72	6	78
Team player	73	4	77
Praise (from manager)	74	0	74
Trains junior/new members of staff	72	0	72
Relationship w/customer	71	0	71
Sickness record	36	25	61
Confident	46	15	61
Good rapport	58	1	59
Learns/takes on new tasks	57	1	58
Efficient	56	1	57
Organized	44	12	56
Works hard/is conscientious	54	0	54
Professional	52	1	53
Enhanced results of area,unit dept.	49	1	50
Punctuality	33	15	48
Supervision/leadership skills	37	9	46
Flexible	44	2	46
Commitment	43	2	45
Experienced	45	0	45
Alert, Aware	38	2	40

Table 13. Positive, Negative, and Total Occurrence of Performance Factors Derived from Content Analysis (Continued)

Table 13. (Continued)

Performance Factor	Positive Occurrences	Negative Occurrences	Total
Adjustment to change	30	7	37
Quick learner	35	0	35
Reliable	33	1	34
Productivity	29	4	33
Attitude	20	8	28
Communications	22	5	27
Stress tolerance/pressure	20	6	26
Positive	24	1	25
Can-do	24	1	25
Calm, Patient	21	3	24
Prioritization	19	5	24
Versatile	21	1	22
Thorough	17	2	19
Personal appearance	17	2	19
Attendance	13	5	18
Responds positively to challenges	15	0	15
Focus, Concentration	3	11	14
Makes suggestions	11	0	11
Mature	7	0	7
Methodical	6	1	7
Able to work unsupervised	7	0	7
Innovative	7	0	7
Honest	6	0	6
Potential for development	5	0	5

Performance Factor	Positive Occurrences	Negative Occurrences	Total
Working with Others (C)			685
Popularity/social skills/polite	116	2	118
Cooperative/helpful	134	1	135
Team player	73	4	77
Friendly/cheerful	99	0	99
Helps out with other tasks	126	0	126
Good rapport	58	1	59
Relationship w/customer	71	0	71
Commitment (C)			177
Works extra long hours early or late	77	1	78
Works hard/is conscientious	54	0	54
Commitment	43	2	45
Trains junior/new members of staff	72	0	72
Cross sell, Sales, Referrals (T)	158	52	210
Execution of Work (T)			482
Accurate	163	43	206
Speed	124	15	139
Organized	44	12	56
Prioritization	19	5	24
Efficient	56	1	57
Productivity (T)			83
Enhanced results of area, unit	49	1	50
Productivity	29	4	33
Motivation (C)			246
Motivated/enthusiastic	72	6	78
Confident	46	15	61
Positive	24	1	25
Alert, Aware	38	2	40
Attitude	20	8	28
Focus, Concentration (C)	3	11	14

Table 14. Performance Factor Clusters (Continued)

Table 14. (Continued)

Performance Factor	Positive Occurrences	Negative Occurrences	Total
Standards			176
High standard/quality	85	4	89
Reliable	33	1	34
Professional	52	1	53
Adaptability (C)			131
Adjustment to change	30	7	37
Versatile	21	1	22
Flexible	44	2	46
Stress tolerance/pressure	20	6	26
Knowledge and Learning New Tasks (C)			235
Quick learner	35	0	35
Knowledgeable/inquisitive	86	11	97
Experienced	45	0	45
Learns/takes on new tasks	57	1	58
Initiative (C)			105
Shows initiative	62	18	80
Makes suggestions	11	0	11
Able to work unsupervised	7	0	7
Innovative	7	0	7
Personal appearance (C)	17	2	19
Customer service (T)	158	7	165
Awards/Recognition			320
Awards/certificates	81	0	81
Praise (from manager)	74	0	74

Note. T=Task performance. C=Contextual performance.

Reliability Analysis

Reliability of coding is important to ensure that coders can reliably share the same codes and that the constructs identified are shared (Ryan & Bernard, 2000). A random sample of 60 forms (just over 10 percent of the sample with completed summary sections) was coded by two independent raters: One I/O student and a psychology professor. Interrater agreement on the coding of this random sample was calculated by computing the percentage of agreement (Ryan & Bernard, 2000) between raters on each appraisal. This was computed as: for each appraisal form, the number of exact matches divided by the total number of codes marked by both coders combined. The average agreement across the 60 forms was then calculated. Interrater agreement was 78 percent, which is considered an acceptable standard (Krippendorf, 1980). The I/O student then coded the remaining forms.

In the present research, the second stage of content analysis, conceptual model building, involved testing the factors against existing theories based on the hypotheses posited. Specifically, the content of cells A to L in Figure 6 will be examined relative to theoretical explanations for and empirical findings on group differences in supervisory performance ratings.

Word Counts: The second general approach to analyzing texts, word counts, can be used to identify ideas or patterns within a text and to provide data for systematic comparisons across groups (Ryan & Bernard, 2000). This method does not consider the context in which the words are used or whether they are used negatively or positively. Thus, it has no evaluative component. It is purely descriptive. Such analyses have been used to identify underlying themes and constructs within texts (Ryan & Bernard, 2000).

Word counts were performed for the main ethnic categories to determine whether there are differences in the number of descriptors or factors mentioned by supervisors in evaluating each group.

CHAPTER 4

RESULTS

The results are presented in two major sections: 1) the comparison of overall and dimensional supervisor ratings in response to Research Questions 1 and 2 and the corresponding hypotheses; and 2) the analysis of the supervisor's narrative summary of performance in response to Research Questions 3 and 4 and those hypotheses.

Analysis of Supervisor Ratings

Hypothesis 1

Hypothesis 1 predicted that minority groups would score lower on overall ratings of performance. As reported in the previous chapter, there are significant differences in tenure across ethnic groups. Ethnic minorities are significantly higher in tenure than majority staff. As reported, tenure is also significantly correlated with overall performance ratings for the majority and black samples. For this reason, an analysis of covariance (ANCOVA) was computed to control for the effect of tenure on the dependent variable (Pedhazur, 1983). Means of supervisors' overall performance ratings for the three appraisal periods were computed using a univariate ANCOVA. Ethnic group was

coded into two categories: majority and minority. Means (shown in Table 15) were examined for significant differences in each appraisal year for the two groups. As shown in Tables 16 to 18, there were significant F ratios for years 1 to 3. Majority staff were rated significantly higher on overall performance than ethnic minority staff. Tenure was a significant covariate (see Tables 16, 17, and 18).

	Majority	sd	Minority	sd	F
Year 1	3.5 (502)	.61	3.2 (107)	.59	14.40**
Year 2	3.5 (524)	.59	3.2 (106)	.47	21.24**
Year 3	3.3 (557)	.55	3.2 (110)	.49	5.07*

Note. Values in parentheses represent n

*p <.05 **p < .01

Table 15. Mean Performance Ratings By Minority/Majority Classification for Each Appraisal Year.

Source	*df*	F	p
Tenure	1	20.78	.00
Ethnic Group	2	14.40	.00
Error	606	(.359)	

Note. Values in parentheses represent mean square errors.

Table 16. Analysis of Covariance for Comparison of Majority/Minority Means -- Year 1

Source	*df*	F	p
Tenure	1	27.20	.00
Ethnic Group	2	21.24	.00
Error	627	(.319)	

Note. Values in parentheses represent mean square errors.

Table 17. Analysis of Covariance for Comparison of Majority/Minority Means – Year 2

Source	*df*	F	p
Tenure	1	25.73	.00
Ethnic Group	2	5.07	.02
Error	664	(.291)	

Note. Values in parentheses represent mean square errors.

Table 18. Analysis of Covariance for Comparison of Majority/Minority Means – Year 3

Given the number of different minority ethnic groups included in the 'minority' classification, another univariate ANCOVA was performed using the racial groups, black, white, and Asian. Levene's test of equality of error variances was computed to determine whether the groups meet the homogeneity of variance assumption of the ANCOVA F test. Levene's test was significant for all 3 years (see Table 19) suggesting the possibility of heterogeneous variance across groups. However, inspection of standard deviations, variances (see Table 20) and box plots for each group suggests relatively similar variances. As shown in Table 20 variances across the three groups are similar. However, the more stringent significance criterion, .01, was applied to minimize the

possibility of bias in year 2 in particular. F values for overall ratings in years 1 and 2 were significant at p<.00. As year 3 did not meet this more stringent criterion, Welch's (F_w) test was used. Results for this test showed significant mean differences across groups (F_w=3.33, df1=2, df2=85.77, p<.04) consistent with the ANCOVA result in which tenure was included as a covariate.

	F	df	p
Year 1 Rating	6.37	2	.00
Year 2 Rating	22.69	2	.00
Year 3 Rating	7.44	2	.00

Table 19. Levene's test of Equality of Error Variances for Asian, Black, and Majority Groups

	White	s^2	Black	s^2	Asian	s^2	F
Year 1	$3.5(502)_a$	.374	$3.2(65)_b$	.345	$3.3(42)_b$	.374	7.30^{**}
Year 2	$3.5(524)_a$	.354	$3.2(65)_b$	.172	$3.2(41)_b$	.312	10.69^{**}
Year 3	$3.3(557)_a$	.313	$3.1(64)_b$	.229	$3.3(46)_{ab}$	.261	3.42^{*}

Note. Means in the same row that do not share common subscripts differ at p < .05 in the Fisher least significant difference test. Values in parentheses represent n.

*p < .05. **p < .01.

Table 20. Mean Performance Ratings By Asian, Black, White Classification for Each Appraisal Year

ANCOVA yielded significant F ratios for years 1 to 3. Tenure was a significant covariate. Fisher's least significant difference test was computed to identify where the significant mean differences among groups exist (Collyer & Enns, 1986). Fisher's test consists of two stages, the first being the test of the omnibus F for the ANCOVA. If F is significant, t-tests are computed for all pairwise comparisons (Keppel, 1982). Analysis of the homogeneity of variance suggests that Fisher's test was appropriate for use. Means are reported in Table 20 and the ANCOVA in Tables 21, 22 and 23.

Source	*df*	F	p
Tenure	1	20.44	.00
Ethnic Group	2	7.30	.00
Error	605	(.359)	

Note. Values in parentheses represent mean square errors.

Table 21. Analysis of Covariance for Comparison of Asian, Black, White Means -- Year 1

Source	*df*	F	p
Tenure	1	26.87	.00
Ethnic Group	2	10.69	.00
Error	626	(.319)	

Note. Values in parentheses represent mean square errors.

Table 22. Analysis of Covariance for Comparison of Asian, Black, White Means – Year 2

Source	*df*	F	p
Tenure	1	25.52	.00
Ethnic Group	2	3.42	.02
Error	663	(.290)	

Note. Values in parentheses represent mean square errors.

Table 23. Analysis of Covariance for Comparison of Asian, Black, White Means – Year 3

Comparison of the majority/minority group means showed that minorities are rated significantly lower than majority group members. Comparison of means between black, white and Asian staff also found significant mean differences. The pattern of results for these 3 groups over the 3 appraisal years was different (see Table 20). For year 3, post hoc analyses revealed that significant differences exist between the black and white groups with no significant differences between Asians and the other two groups. That is, black staff are rated significantly lower than white staff on overall performance. There were no significant differences between Asian and white, or Asian and black staff in overall performance. For years 1 and 2, the 2 rating periods prior to year 3, black and Asian staff's overall performance ratings were significantly lower than those of white staff. This finding of lower overall performance ratings for black versus white employees is consistent with research conducted in the United States (e.g. Ford et al., 1986; Sackett & DuBois, 1991; Waldman & Avolio, 1991). The lower rating of black and Asian staff relative to majority employees is also consistent with findings of preliminary research conducted in the United Kingdom (Wilson, 1995).

It has been argued that score (or rating) variance can be studied as an effect in its own right (Collyer & Enns, 1986) as heterogeneity in variance can result from floor or ceiling effects in the measuring instrument. Inspection of variances for the 3 groups in each rating period revealed differences in year 2. Assessment of similarity was made using Collyer and Enns' (1986) criterion: the largest variance should generally not be more than twice the size of the smallest. There were differences in black and Asian variances despite no differences in ratings. The variance in black ratings was almost half that of the other two groups (Asian=.312, black=.172, majority=.354. Year 3 analyses revealed relatively similar variances for the two minority groups (Asian=.261, black=.229, majority=.313). There were no differences in year 1. Black ratings tended to cluster around the center of the rating scale which may reflect possible floor and ceiling effects for this group in years 2 and 3. The remaining analyses will focus on the more detailed comparisons using the white, black, and Asian categories.

In summary, these analyses provide support for Hypothesis 1 which predicted that minority staff would have significantly lower overall performance ratings than majority staff.

Hypothesis 2a

Hypothesis 2a predicted that minority groups would be rated the same on skill areas (dimensional performance) while their scores on overall performance will be significantly different. As described in the previous chapter, using skills ratings for year 3, difference scores were computed on each skill area between level of skill required and level achieved. Assessments of 'level required' and 'level achieved' were made on a 7-point scale where 1 was high and 7 was low. Difference scores for staff who achieved

less than the level required would be negative, while differences for those who achieved higher than the level required were positive. Group differences were examined by comparing mean difference scores across groups using a univariate ANCOVA of tenure. Table 24 shows the results of this analysis.

Dimension	White[a]	Black[b]	Asian[c]	F
Business Awareness	$-.43_a$	-1.00_b	$-.64_{ab}$	3.801*
Business Development	-.08	-.64	-.07	1.980
Communication – Verbal	-.08	-.60	-.03	2.179
Communication – Written	.33	.16	.35	.462
Customer Service	-.03	-.32	.42	2.140
Decision Making	-.41	-.88	-.28	1.844
Initiative	-.57	-1.20	-.57	2.472
Numeracy	.32	.20	.14	.514
Planning & Organizing	.48	.04	.57	1.084
Team Work & Team Leadership	.26	.16	.60	1.038
Use of Systems	$.58_a$	$.08_b$	$.46_{ab}$	3.150*
Work Standards	$.15_a$	$-.52_b$	$.53_a$	3.931*

Note. Means in the same row that do not share common subscripts differ at $p < .05$ in the Fisher least significant difference test.

[a]n = 220. [b]n = 25. [c]n = 28.

*p < .05.

Table 24. Mean Difference between Achieved Performance and Required Performance on Skill (Dimensional) Ratings of Performance.

There were significant F ratios for only three of the 12 skill areas (dimensions) rated – Business Awareness, Use of Systems, and Work Standards. Tenure was not a significant covariate on any of these dimensions but was a significant covariate on two of the 12 dimensions (customer service, and planning and organizing). Levene's test for equality of error variances was computed for each dimension. The F value was

significant for two dimensions, Business Awareness and Use of Systems (F=4.421, df=2, p<.01; and F=3.605, df=2, p<.05, respectively) and non-significant for the remaining 10 dimensions including Work Standards (F=.190, df=2, ns). Inspection of variances for Business Awareness (white=1.02, black=1.41, Asian=1.79) and Use of Systems (white=1.03, black=.74, Asian=.62) did not suggest dissimilar variances across groups. Post hoc analyses, Fisher's least significant difference test, showed that Black staff were rated as having significantly lower achievement than white staff on Business Awareness and Use of Systems and significantly lower achievement than white staff and Asian staff on Work Standards. Further, although there were no significant differences on the remaining nine dimensions, examination of Table 24 shows that the average ratings for black staff are lower than those for majority staff on all 12 dimensions. Using the sign test, the chances of this occurring by chance is less than *p*= .000. No clear pattern emerged for Asian staff vis â vis the majority group.

In summary, there was no support for Hypothesis 2a.

In addressing Hypotheses 2b and 2c, Pearson Product Moment Correlations were computed among the 12 dimensional ratings and overall performance rating for the three ethnic groups. Due to the disparity in sample sizes across the three groups, correlations were computed for a random sample of 27 white staff. The sample size of 27 was chosen to make the 3 groups comparatively equal in size. The Asian sample was n=28 and the black sample n=25. Pairwise deletion of missing data brought the sample sizes to 26 and 22, respectively for Hypothesis 2b. The number 27 was chosen as it fell between 25 and 28. The sample was chosen using the statistical software SPSS's random sample generator. Pairwise deletion of missing values brought this sample size to 23. Due to

pairwise deletion, these figures varied slightly in later analyses depending on the specific procedure. Sample sizes for each analysis are shown in the respective tables. Correlations were also computed for the total sample of white staff for which skills data were available (n=220) to allow comparisons between findings for the random sample and the full sample for this group.

Hypothesis 2b

Hypothesis 2b predicted that there would be a different pattern of relationships among skill (dimensional) and overall performance ratings for majority than minority staff. Intercorrelations were computed between dimensional ratings and overall performance rating for each group. These analyses are presented in Table 25.

Examination of zero-order correlations for both the random sub-sample and full sample of majority staff shows that relative to minority staff, majority staff had more significant correlations between skill ratings and overall performance rating. Specifically, 8 of the skill ratings were significantly correlated with the overall performance rating in the random sub-sample compared to 1 significant correlation for black staff and no significant correlations for Asian staff, providing support for Hypothesis 2b. Ten of the correlations were significant in the full sample of majority staff. Support for this hypothesis is consistent with previous research (e.g. Cascio & Valenzi, 1978; Kraiger & Ford, 1990) and suggests that these ratings may have different underlying meanings across the studied groups.

	Overall Performance Rating			
	Asian[a]	Black[b]	White[c]	White[d]
Dimension	Overall Rating	Overall Rating	Overall Rating	Overall Rating
Business Awareness	.12	.22	.50**	.36**
Business Development	.12	-.13	.43*	.12
Communication – Verbal	-.14	.21	.41*	.25*
Communication – Written	-.06	.03	.31	.24**
Customer Service	.04	.18	.61**	.38**
Decision Making	.01	.35	.49**	.37**
Initiative	.13	.26	.39*	.26*
Numeracy	-.01	-.05	-.06	.04
Planning & Organizing	-.03	.27	.19	.27**
Team Work & Team Leadership	.01	-.18	.55**	.32**
Use of Systems	.02	.16	.18	.28**
Work Standards	.04	.50**	.83**	.45**

Note. [a]n=26. [b]n=22. [c]Random Sample, n=23. [d]n=203-218

*$p < .05$. **$p < .01$.

Table 25. Correlations Between Dimension Ratings (Mean Difference Between Achieved and Required Performance) and Overall Performance Rating.

To further explore the relationship between the dimensional and overall ratings, stepwise regression of dimension ratings on the overall rating was computed for each ethnic group. In this procedure, tests are performed at each step of the analysis to determine the contribution of each predictor already in the equation if it were to be

entered last. Predictors are eliminated from the equation if they have lost their usefulness upon introduction of new predictors (Pedhazur, 1982). Results of these analyses are shown in Tables 26 and 27.

As shown in Table 26, four dimensions predict overall rating for majority staff, Work Standards, Business Awareness, Use of Systems, and Customer Service. Wilkinson's (1979) tables were used to test the significance of R^2. Wilkinson (1979) has

Variable	β	B	t	p	R^2	ΔR^2
Step 1						
Work Standards	.191	.46	7.45	.00		
F(1,200)=54.12**					.21	
Step 2						
Work Standards	.153	.37	5.63	.00		
Business Awareness	.13	.23	3.57	.00		
F(2,199)=35.05**					.26	.04
Step3						
Work Standards	.14	.34	5.26	.00		
Business Awareness	.12	.21	3.23	.00		
Use of Systems	9.248E-02	.17	2.74	.00		
F(3,198)=26.64**					.28	.02
Step 4						
Work Standards	.11	.27	3.68	.00		
Business Awareness	.10	.18	2.74	.00		
Use of Systems	9.368E-02	.17	2.80	.00		
Customer Service	6.527E-02	.15	2.12	.03		
F(4,197)=21.46**					.30	.01

Note. Dependent variable: Overall Rating. N=202. **P<.01.

Table 26. Stepwise Regression Predicting Overall Ratings from Dimensional Ratings (Majority Group)

Variable	β	B	t	p	R^2	ΔR^2
Step 1						
Work Standards	.204	.50	2.77	.01		
$F(1,22)=7.69^{**}$					.25	

Note. Dependent variable: Overall Rating.

N=24. $^{**}P<.01$.

Table 27. Stepwise Regression Predicting Overall Ratings from Dimensional Ratings (Black Staff)

criticized researchers for using inflated R^2 statistics in reporting results of stepwise regression analyses. He points out that the significance tests reported in widely used computer programs do not test the appropriate *F*; that this statistic is printed at each step although the *F* distribution is unknown under automated stepwise selection (Wilkinson, 1979). Using *monte carlo* methods, he provides significance tables which he recommends that users of automated stepwise computer programs consult to evaluate the significance of the final regression equation they select. These tables are more conservative than the usual *F* tables (Myers & Wells, 2003). This model accounted for 30 percent of the variance in overall performance rating. Using Wilkinson's tables, R^2 was significant. R^2 (k,m,n,α)=.13; where k=number of predictors selected; m=number of predictors; n=sample size; and α=alpha level. Per Wilkinson's tables, R^2 (4,15,202,.01)=.13 Thus R^2 must exceed .13 to reject the null hypothesis at a critical value of .01.

In contrast to the findings for the majority group, only one dimension, Work Standards, predicted overall rating for black staff in a stepwise regression analysis (see Table 27). This model accounted for 25 percent of the variance in overall rating. R^2 was

not significant. Using Wilkinson's tables, R^2 (1,15,24,.01)= .36 and R^2 (1,15,24,.05=.29). R^2 must exceed .36 to reject the null hypothesis at an alpha level of .01 and exceed .29 to reject the null hypothesis at a critical value of .05.

There was concern regarding the stability of the model obtained for black staff given the susceptibility of this procedure to capitalization on chance, particularly with the use of small samples (Myers & Wells, 2003). Another concern was the large correlation between Work Standards and overall performance ($r_{ws.overall}$=.50) given the small and near-zero correlations between overall rating and the other dimensions. For these reasons, a simultaneous regression analysis was also performed for this group (see Table 28).

	B	t
Business Awareness	.45	2.12*
Business Development	-.06	-.23
Communication – Verbal	-.14	-.66
Communication – Written	-.43	-2.27*
Customer Service	.28	1.38
Decision Making	-.64	-1.40
Initiative	.95	2.54*
Numeracy	-.45	-1.71
Planning & Organizing	.30	.88
Team Work & Team Leadership	-.54	-2.63*
Use of Systems	.44	1.83
Work Standards	.68	1.74
R^2=.77, *F*=3.06*		
Adjusted R^2=.51		

Note. Dependent variable: Overall Rating.

N=24. *P<.05

Table 28. Simultaneous Regression Analyses Predicting Overall Ratings from Dimensional Ratings (Black Staff)

The results of the simultaneous analysis show that four dimensions, Business Awareness, Written Communication, Initiative, and Team Work and Team Leadership predict overall performance rating for black staff. Fifty-one percent of the variance in overall rating was accounted for by this model. The effect size was .71. An effect size above .5 can be considered a large effect (Cohen, 1988). This different finding with the use of a simultaneous analysis suggests that although 11 of the 12 dimensions did not meet the criteria for inclusion in the stepwise regression equation, when using simultaneous entry, four of these variables make a significant contribution over and above the contribution of the others to the prediction of the dependent variable. A simultaneous regression analysis for the majority group supported the stepwise findings for that group – Business Awareness, Customer Service, Use of Systems, and Work Standards were the only significant predictors of overall performance. This model accounted for 27 percent of the explained variance. The effect size was .51.

Regression analyses for the Asian sample produced contrasting results from those for the other two groups. In the stepwise analysis, none of the dimensions met the criterion for inclusion in the model (significance of .05), thus no variables were entered in the equation. In short, none of the dimensions predicted overall rating. A simultaneous analysis supported this finding.

In summary, these regression analyses show that the relationship between the dimensional and overall ratings differed among the three ethnic groups, providing further support for Hypothesis 2b.

Hypothesis 2c

Hypothesis 2c predicted that there would be less dimensionality among skill ratings for majority than for minority staff. Intercorrelations among dimensions were computed with the effect of tenure and overall performance rating partialled out. The objective was to identify the relationship among the skill areas for the three groups at the same level of performance. These data are shown in Tables 29, 30 and 31.

Dimen.	2	3	4	5	6	7	8	9	10	11	12
1. BusA	51**	26	07	39	27	46*	-03	45*	27	-16	41*
2. BusDevelop		03	00	26	-01	26	-11	28	-02	06	07
3. Comm-Verbal			35	74**	22	38	-29	45	46*	-23	61**
4. Comm-Written				23	26	45*	31	13	17	-02	55**
5. Customer Service					26	29	52**	32	48**	-29	59**
6. Decision Making						65**	14	32	30	-13	18
7. Initiative							20	48**	42*	-26	59**
8. Numeracy								-04	-03	25	-09
9. Planning and Organizing									57**	-43*	74**
10. Team Work and Team Leadership										-38	66**
11. Use of Systems											-62**
12. Work Standards											

Note. n=21.

*$p < .05$.

**$p < .01$.

Table 29. Partial Correlations Among Dimension Ratings with Tenure and Overall Rating Partialled Out – Random Sub-Sample of Majority Staff

Dimen.	2	3	4	5	6	7	8	9	10	11	12
1. BusA	26	27	24	-03	50**	32	46*	03	11	16	05
2. Bus Develop		21	11	-21	58**	76**	50**	21	05	23	-21
3. Comm-Verbal			-09	30	42**	28	03	18	27	11	10
4. Comm-Written				-04	26	15	32	12	-24	39	28
5. Customer Service					-05	-07	-04	-03	31	-04	18
6. Decision Making						74**	53**	41*	12	11	32
7. Initiative							54**	05	03	16	-25
8. Numeracy								42*	-23	28	-01
9. Planning and Organizing									-19	-24	52**
10. Team Work and Team Leadership										26	28
11. Use of Systems											04
12. Work Standards											

Note. n=*20.*

*$p < .05$.

**$p < .01$.

Table 30. . Partial Correlations Among Dimension Ratings with Tenure and Overall Rating Partialled Out – Black Staff.

Dimen.	2	3	4	5	6	7	8	9	10	11	12
1. BusA	55**	40*	30	38*	54**	55**	49**	15	18	16	28
2. BusDevelop		19	46**	04	47**	56**	45**	12	14	33	01
3. Comm-Verbal			08	54**	60**	44**	18	49**	57**	-20	35
4. Comm-Written				06	48**	32	25	24	35	21	28
5. Customer Service					54**	48**	37	17	59**	28	30
6. Decision Making						75**	54**	36	54**	11	41**
7. Initiative							52**	38**	53**	11	55**
8. Numeracy								21	34	26	32
9. Planning and Organizing									42*	-19	60**
10. Team Work and Team Leadership										03	50**
11. Use of Systems											-28
12. Work Standards											

Note. n=*26.*

*$p < .05$.

**$p < .01$.

Table 31. Partial Correlations Among Dimension Ratings with Tenure and Overall Rating Partialled Out -- Asian Staff.

Examination of the number of significant correlations across groups shows that there were more significant intercorrelations among dimensions for the Asian sample compared to the other two groups. Black staff had the fewest number of significant correlations. Specifically, 30 of the 66 intercorrelations were significant for the Asian sample, compared to 21 for the majority group, and 11 for the Black sample. Hypothesis

2c was not supported. However, the pattern of intercorrelations among dimensions was different across groups. For example, Business Development was not significantly correlated with any other skill areas for the majority sample, but is highly and significantly correlated with Decision Making (.58, p<01), Initiative (.76, p<.01) and Numeracy (.50, p<.01) for Black staff; and Verbal Communication (.46, p<01), Decision Making (.47, p<.01), Initiative (.56, p<.01), and Numeracy (.45, p<.01) for Asian staff. The corresponding correlations are non-significant and close to zero for majority staff. Similarly, Work Standards is highly and significantly correlated with 8 skill areas for majority staff (Business Awareness, Verbal Communication, Written Communication, Customer Service, Decision Making, Initiative, Planning and Organizing, Team Work and Team Leadership, and Use of Systems) whereas Work Standards is significantly correlated with only one other dimension for Black staff (Planning and Organizing) and 4 dimensions for Asian staff (Decision Making, Initiative, Planning and Organizing, and Team Work and Team Leadership).

In summary, Hypothesis 2c predicted less dimensionality among skill ratings for majority than minority staff. This hypothesis was not supported as there were more intercorrelations among dimensions for one minority group, Asian staff, rather than for the majority sample.

Analysis of Supervisor's Narrative Summary

Hypothesis 3a

Hypothesis 3a predicted differences in the factors cited by supervisors in their written justifications of their evaluation. Specifically, that there would be proportionally

fewer positive mentions of task factors; and more mean negative mentions of task factors for minorities compared to majority group members.

To test this hypothesis, the mean occurrences of positive and negative mention of task factors were computed for each group; that is, the average number of times that supervisors mentioned task factors in a positive and negative manner for each group. A univariate ANCOVA was computed to test for significant mean differences across groups while controlling for the effect of overall performance rating on the dependent variable. These factors were compared in the subjectively derived clusters. Mean occurrence was computed for 'Sales, Cross Sales, Referrals'; 'Productivity'; 'Knowledge and Learning New Tasks'. Support for this hypothesis would be reflected by significantly higher means on positive mention for white staff compared to the minority groups and higher means on negative mention for minority compared to majority staff. Results of these analyses are shown in Table 32.

For each of the positive task factors, overall performance rating was a significant covariate. Overall rating was a significant covariate for only two negative factors, Execution of Work, and Sales. Specific results of the ANCOVA are reported as Appendix D. Levene's test for equality of error variances was computed to test for homogeneity of variance across groups. This test was significant for all positive factors except Productivity; however, inspection of the variances across groups does not suggest large differences. These results are also reported in Appendix D. This test was also positive for two of the negative factors, Knowledge and Learning New Tasks, and Sales. To minimize the possibility of bias, an alpha level of .01 was used as the significance test for *F*.

Task Factors	Mean Occurrence			F
	Asian[1]	Black[2]	White[3]	
Positive Mention				
Sales, cross sales, referrals	$.30_a$	$.25_b$	$.18_b$	3.44^{**}
Productivity	.04	.06	.08	.378
Knowledge & Learning New Tasks	.43	.20	.28	2.79
Execution of Work	1.02_a	$.44_b$	$.57_b$	4.66^{**}
Negative Mention				
Sales, cross sales, referrals	.04	.11	.06	.914
Productivity	.00	.00	.00	.00
Knowledge & Learning New Tasks	.02	.03	.00	1.36
Execution of Work	.13	.10	.08	.539
Positive and Negative Combined				
Sales, cross sales, referrals	$.34_{ab}$	$.35_a$	$.24_b$	3.29^{*}
Productivity	.04	.06	.07	.378
Knowledge & Learning New Tasks	.43	.18	.27	2.49
Execution of Work	$.95_a$	$.43_b$	$.55_b$	4.95^{**}

Note. Means in the same row that do not share common subscripts differ at $p < .05$ in the Fisher least significant difference test. [1]n=46. [2]n=64. [3]n=557.

*p<.05. **p<.01.

Table 32. Mean Occurrence of Positive and Negative Mention of Task Factors by Group.

There were significant group differences on only two positive task performance factors, 'Sales, Cross Sales, Referrals', and 'Execution of Work'. The direction of the differences was the opposite of that predicted. Asians were significantly higher than majority and black staff on these factors. In other words, irrespective of level of performance (as measured by overall performance rating), supervisors were more likely to positively mention 'Sales', and how work is executed (e.g. accuracy, speed, organization) in justifying their ratings of Asian staff. There were no significant differences across groups in the negative mention of task factors. These results reflect the low frequency of negative factors overall. There were only 5 negative comments regarding Productivity and low frequencies on the other factors (as shown in the previous chapter).

In effect, sales and execution of work were more salient to supervisors in rating Asians compared to majority and black staff.

Hypothesis 3a was not supported. However, the general principle -- that there are differences in the factors cited by supervisors in their written justifications of their evaluations irrespective of performance level, is supported. Certain performance factors are more salient to supervisors in rating one subgroup compared to another.

Hypothesis 3b

Hypothesis 3b predicted that supervisors would emphasize contextual elements of performance more for minority than majority staff in justifying their performance evaluations. Support for this hypothesis would be reflected by higher mean occurrences of these factors for minority staff than majority staff. Means were compared on contextual performance factors cited by supervisors. Means were computed for both

positive and negative occurrences of each factor combined. The interest here was not whether the comment was positive or negative, but the salience of contextual factors to supervisors as they justified their evaluations. An ANCOVA was computed with overall performance rating as the covariate. Levene's test for equality of error variances across groups was computed for each omnibus F for the ANCOVA. This test was significant for Adaptability ($F=12.801$, $df=2$, $p<.00$) and Standards ($F=3.813$, $df=2$, $p<.02$). Results of the mean comparisons are presented in Table 33.

	Mean Occurrence			
Contextual Factors	Asian[1]	Black[2]	White[3]	F
Working with Others	1.15	1.00	.88	1.41
Motivated/Enthusiastic	.24	.17	.19	.330
Adaptability	$.33_a$	$.13_b$	$.15_b$	4.6^{**}
Commitment	.50	.38	.40	.43
Standards	.20	.20	.29	.78
Initiative	.06	.10	.09	.488

Note. Means in the same row that do not share common subscripts differ at $p < .05$ in the Fisher least significant difference test.

[1]n=46.

[2]n=64.

[3]n=557.

**p<.01.

Table 33. Mean Occurrence of Contextual Performance Factors by Group.

There were significant group differences on only one of the contextual factors examined, 'Adaptability'. The .01 criterion was used in testing the significance of *F*. Overall performance rating was a significant covariate for this dependent variable (F=3.878, df=1, p<.04).

The significant group difference was in the predicted direction, however, there was a significant finding for only one of the two minority groups and for only one cluster of factors. This difference is consistent with that for Hypothesis 3a. Adaptability was cited significantly more by supervisors in describing Asian staff compared to Black and majority staff at the same level of performance. Examination of the 'sign' of the comment revealed that the descriptions of Asian staff on this factor were positive. Hypothesis 3b received minimal support.

Hypothesis 4a

Hypothesis 4a predicted that negative performance factors such as attendance, punctuality, and accuracy will be more salient to supervisors in justifying minority evaluations compared to majority staff evaluations. As described previously, these are factors that are considered negative indices of performance (regardless of whether they are mentioned in a positive way). This hypothesis suggests that supervisors will focus on negative indices of performance by mentioning these factors (either positively or negatively) in justifying ratings of minority group members. Groups were compared on mean occurrences of attendance, punctuality, accuracy, and sickness record. Means were computed for total occurrence: positive and negative mention of each factor combined as the interest here is in the salience of these factors to supervisors in summarizing performance.

A univariate ANCOVA was computed to compare group means. Overall performance rating was included as the covariate. Again, Fisher's test was used as the post hoc test where the F for the ANCOVA was significant. Levene's test for homogeneity of variance across groups was significant for one factor only, Accurate (F=12.743, df=2, p<.00). Inspection of variances suggests higher variance within the Asian sample relative to the black sample. The results of this analysis are shown in Table 34.

	Mean Occurrence			
Performance Factors	Asian[1]	Black[2]	White[3]	F
Accurate	$.43_a$	$.17_b$	$.23_b$	5.38**
Punctuality	.04	.04	.05	.109
Attendance	.02	.01	.01	.04
Sickness Record	.04	.09	.07	.52

Note. Means in the same row that do not share common subscripts differ at p < .05 in the Fisher least significant difference test.

[1]n=46. [2]n=64. [3]n=557. **p<.00.

Table 34. Means on Negative Indices of Performance by Group.

One of the four factors, Accuracy, was significantly different across the 3 groups. Asian staff have the highest mean occurrence on this factor. Overall performance rating was not a significant covariate on Accuracy (F=1.823, df=1, n.s.) or on any of the dependent variables. Fisher's least significant difference test showed that mean occurrence of this factor was significantly higher for Asian staff relative to majority and black staff. Given the possibility of heterogeneous variances, Welch's (F_w) test was

computed. This test was significant (F_w=3.036, df1=2, df2=97.06, p<.05). In interpreting the Welch test results, it is important to note that this test does not allow a covariate, thus the effect of overall performance on the dependent variable has not been controlled. However, as overall performance was not a significant covariate, the Welch test result should be generally equivalent to the ANCOVA.

Essentially, supervisors were significantly more likely to mention Accuracy in describing the performance of Asian staff compared to black and majority staff. This result was at the same level of performance across groups. Inspection of the frequency of negative versus positive comments regarding Accuracy shows that these comments were positive. The non-significant finding for Attendance and Punctuality may have been a result of the low occurrence of these factors. Supervisors did not generally mention Attendance or Punctuality in justifying their performance ratings. In summary, Hypothesis 4a received partial support.

Hypothesis 4b

Hypothesis 4b predicted differences across groups, at the same level of performance, in supervisors' tendency to make positive comments in their written justifications of performance. In a test of this hypothesis, the total mean occurrence of positive factors for each group was computed. A univariate ANCOVA was performed to test for significant differences across groups while controlling for the effect of overall performance rating on the dependent variable. Levene's test was not significant, suggesting homogeneity of variance across groups. Table 35 shows the results of the mean comparisons.

Performance Factors	Mean Occurrence			
	Asian[1]	Black[2]	White[3]	F
Positive Comments	4.86_a	3.04_b	3.36_b	3.893*

Note. Means in the same row that do not share common subscripts differ at $p < .01$ in the Fisher least significant difference test.

[1]n=46. [2]n=64. [3]n=557.

*p<.05.

Table 35. Total Mean Occurrence of Positive Comments by Group.

Overall performance rating was a significant covariate (F=21.663, df=1, p<.01). There were significant differences across groups in positive comments made by supervisors. Asians were more likely to receive positive comments from supervisors than majority and Black staff at the same level of performance. Hypothesis 4b was supported.

Hypothesis 4c

Hypothesis 4c predicted that supervisors would emphasize liking factors more for majority staff than minority staff in their written summaries of performance. Support for this hypothesis would be indicated by higher mean occurrences of positive mention of positive affect or liking factors for majority staff relative to minorities at the same level of performance. The three ethnic groups were compared on three performance factors: Friendly/Cheerful, Positive, and Praise from supervisor. A univariate ANCOVA was computed with overall performance rating as a covariate. Fisher's test was used as the post hoc test to determine specifically where mean differences exist. Results of the mean comparisons are reported in Table 36.

Performance Factors	Mean Occurrence			
	Asian[1]	Black[2]	White[3]	F
Friendly/Cheerful	.21	.14	.11	2.177
Positive	$.08_a$	$.01_b$	$.01_b$	4.715**
Praise from Manager	$.24_a$	$.11_b$	$.07_b$	8.536**

Note. Means in the same row that do not share common subscripts differ at p < .00 in the Fisher least significant difference test

[1]n=46. [2]n=64. [3]n=557.

**p<.01.

Table 36. Mean Occurrence of Positive Affect/Liking Factors by Group.

Levene's test was significant for all three factors. Examination of variances and box plots show more variance within the Asian sample relative to the other two groups and similar variances for the majority and black samples. The stricter significance criterion of p<.01 was used in testing F for the ANCOVA. Overall performance rating was a significant covariate on one dependent variable, Praise from Manager (*F*=6.521, df=1, p<.01). Welch's test was also computed given possible heterogeneity of variance. Again, this test does not control for the effect of overall performance. Results are shown in Table 37. Means are significantly different for only one factor, Praise from Manager.

Using Welch's test, there were significant differences across groups on only one of the three liking factors examined, Praise from Manager. Also, the findings are the converse of those predicted. One minority group, rather than majority staff, was significantly higher on these factors. Managers were more likely to praise Asian staff in

Performance Factors	F_w	df1	df2	*p*
Friendly/Cheerful	1.111	2	93.316	.334
Positive	1.719	2	90.082	.185
Praise from Manager	3.168	2	91.448	.047

Note. Majority n=635. Asian n=54. Black n=76.

Table 37. Welch's test of equality of means for Positive Affect/Liking Factors by Group

their written justifications of performance compared to majority and black staff. This finding is consistent with those for Hypotheses 3a, 3b, 4a and 4b which show higher mean occurrences for Asian staff on the examined factors relative to majority and black staff at the same level of performance. Although group differences were found at the same level of performance, Hypothesis 4c was not supported given the direction of these differences.

General Research Question

Finally, a general research question addressed the broad nature of the comments made by supervisors across groups. The interest here was simply what supervisors were generally saying in justifying their ratings of different groups. Although mean differences in factors have been presented earlier, this analysis presents a visual picture of supervisors' comments across groups at different categories of performance (using the framework presented in Figure 6). Relevant questions would be: are there differences in how much supervisors write across groups? Is there a broad range of factors listed in each group or simply a few salient factors? How consistently do supervisors mention each given factor within each group (i.e. what percentage of staff receive particular comments within a given group)? Frequencies of comments were computed for different

levels of overall performance: good performance (ratings of 4 and above), average performance (ratings of 3), and poor performance (ratings of 2 and below). Figures 7, 8 and 9 show the percentage of staff in each group for whom specific factors were mentioned at each category of performance. The factors in bold are those that are mentioned in more than one group. Contingency coefficients were computed for the total group sample on each factor rather than the sample at each performance level due to the small sample sizes. For example, a significant coefficient on Accurate, reflects a significant difference in distributions across all performance levels. Sample sizes for high and poor performers were too low to allow any meaningful interpretation. However, examination of results for average performers gives an indication of the factors salient to supervisors in rating different groups. For good performers, supervisors focused primarily on sales (task performance) and helping behaviors across all groups. This emphasis was strongest for black staff – almost half were described as performing well in terms of cross selling, sales, and referrals; and almost 40 percent were described as helping with other tasks. This suggests that sales is a particularly salient factor in rating these staff. As the sample sizes for the black and Asian groups were very small this finding is not conclusive but suggests a possible difference in the salience of factors across groups.

One clear finding for average performers is that supervisors wrote considerably more in describing the performance of Asian staff compared to the other two groups at the same level. Their map of performance for this group was more variable in terms of range of factors mentioned. At the same time, however, there was more consistency in their comments. That is, percentages were generally higher than for the other two

	Performance Factors	%
Asian Ratees (n=13)	**cross sell, sales, referrals**	31
	cooperative/helpful	23
	speed	23
	customer service	23
Black Ratees (n=13)	**cross sell, sales, referrals**	46
	helps out with other tasks	39
	team player	31
	cooperative/helpful	23
	customer service	23
	supervision/leadership	23
	initiative, proactive	23
Majority Ratees (n=179)	**helps out with other tasks**	29
	cross sell, sales, referrals	27
	customer service	27
	cooperative/helpful	20
	high standard/quality	20
	speed	17
	knowledgeable, inquisitive	16
	initiative, proactive	16

Note. [a]Frequencies for negative factors were too low to allow inclusion.

Figure 7. Positive Factors Mentioned for Good Performers by Group.[a]

groups. For example, 42 percent of Asians were described as accurate compared to 15 percent of black and 17 percent of majority staff. As discussed previously, frequencies for negative factors were low. Accuracy and sales were mentioned most frequently as negative factors. Sales was mentioned relatively more frequently for black staff consistent with its relative importance for good performers above.

Findings for poor performers are shown in Figure 8 however, sample sizes are too small for analysis. There were no Asian staff in this category.

Performance Factors

	Positive Factors	%	Negative Factors	%
Asian	**accurate***	42	**accurate**	9
(n=33)	**cross sell, sales, referrals**	30	focus, concentration	6
	customer service	30		
	praise (from manager)*	27		
	cheerful, friendly	24		
	Speed	24		
	learns/takes on new tasks	21		
	cooperative/helpful	18		
	popular/social skills/polite	18		
	awards/certificates	18		
	helps out with other tasks	18		
	good rapport	18		
	experienced*	15		
	commitment*	15		
Black	**cross sell, sales, referrals**	21	**cross sell, sales, referrals**	13
(n=48)	**cooperative/helpful**	21	initiative, proactive	6
	popular/social skills/polite	19		
	accurate	15		
	cheerful, friendly	15		
Majority	**accurate**	17	**cross sell, sales, referrals**	8
(n=367)	**customer service**	15	**accurate**	5
	cross sell, sales, referrals	14		

Note. Distributions significantly different at $p<.05$.

Figure 8. Positive and Negative Factors Mentioned for Average Performers by Group.

	Performance Factors			
	Positive Factors	%	Negative Factors	%
Asian (n=0)				
Black (n=3)	good rapport	33	**accurate**	33
	personal appearance*	33	punctuality	33
	alert, aware	33	cross sell, sales, referrals	33
	knowledgeable, inquisitive	33	attitude	33
			supervision/leadership	33
			initiative, proactive	33
Majority (n=11)	popular/social skills/polite	27	**accurate**	36
	accurate	18	motivated/enthusiastic	18
	confident	18	customer service	18
	awards/certificates	18		

Note. Distributions significantly different at $p<.05$.

Figure 9. Positive and Negative Factors Mentioned for Poor Performers by Group.

CHAPTER 5

DISCUSSION

This study represents an empirical attempt to identify differences in the underlying meaning of job performance ratings across groups. Differences between majority and minority group performance ratings have been attributed to psychological and perceptual biases on the part of the rater (e.g. Ilgen & Youtz, 1990); and differences in ability (e.g. Schmidt & Hunter, 1981) or experiences at work (e.g. Ilgen & Youtz, 1990). Although there is some research that has identified race effects in ratings (Kraiger & Ford, 1985), there is very little research aimed at exploring the processes underlying minority group differences in ratings. Specifically, researchers have not sought to look beyond ratings and explicate the performance construct across groups.

This study used a differential constructs approach to explore the extent to which the underlying dimensions of ratings might vary by racial group; thereby addressing the paucity of research in this area.

Overview of Results

The first research question suggested that consistent with existing research on minority group differences in performance ratings, there would be significant differences in performance ratings between minority and majority group members. Hypothesis 1

predicted that minorities would be lower on supervisors' overall ratings of performance than majority group members. This hypothesis was supported. Majority group members had significantly higher ratings than the minority group for the 3 years examined. When ratings were compared for Asian, black, and white classifications, white staff had significantly higher ratings than Asian and black staff in years 1 and 2. White staff had significantly higher ratings than black staff in year 3 while there were no significant differences between Asian staff and the other two categories. The finding of significantly lower performance ratings for minority staff is consistent with previous U. S. research (Ford et al., 1986). The differences between Asian, black, and white ratings is consistent with preliminary research in the U. K. In a sample of 3 public sector organizations, Wilson (1995) found that Asian and black staff received significantly lower overall performance ratings than their majority counterparts and were significantly more likely to be rated not suitable for promotion.

Research Question 2 addressed the relationships among overall ratings and dimensional ratings. Overall performance ratings for year 3 (the most recent provided) were used in this and all subsequent analysis. Hypothesis 2a predicted that the groups would be rated the same on skills while their overall performance ratings would be significantly different. This hypothesis was not supported. There were significant differences between majority and black staff in overall performance and significant differences across these groups on the dimensional ratings, as shown by the sign test There were no significant differences between Asian staff and the majority group in either dimensional ratings or overall ratings (for year 3). There were also no significant

differences between Asian and black staff in overall ratings and no differences on 11 of the 12 dimensional ratings.

Hypothesis 2b predicted that skill and overall ratings would show a different pattern of relationships within the majority sample compared to the minority groups. This hypothesis was well supported by correlational and regression analyses. There are more statistically significant zero-order correlations between overall performance and dimensional ratings for the majority sample compared to the two minority groups. In addition, the regression analyses suggest that overall performance ratings have a different underlying meaning across the three ethnic groups; the set of dimensional ratings that predicted the overall ratings was different for the black employees compared to the majority employees, and none of the dimensional ratings predicted the overall ratings for the Asian employees. What these differences mean is not entirely clear. The absence of a relationship between dimensional and overall measures for the Asian group suggests that the overall rating may simply represent a global impression or factors other than those included on the rating form are operating. Although four dimensions predicted overall performance ratings for majority and black staff, there was overlap on only one of the four factors – Business Awareness. One may speculate that the significant predictors for black staff (Business Awareness, Written Communication, Initiative, and Team Work and Team Leadership) are areas in which supervisors believe that black staff are traditionally poor. For example, in one UK study (Wilson, 1995) staff reported that they were told by supervisors that black staff cannot write therefore they were not eligible for senior administrative positions. Similarly, black staff may be perceived as demonstrating less initiative and perhaps considered less likely to work cooperatively in teams.

Supervisors may possess different beliefs regarding majority staff, or may not have the same concerns regarding these particular factors, hence the use of different predictors. This is mere speculation and represents an area for future research. Taken together, the correlational and regression analyses suggest that overall performance ratings might not be comparable across groups and may reflect different underlying meanings.

Hypothesis 2c predicted less dimensionality among dimension ratings for the majority group compared to the minority staff. This hypothesis was not supported. In fact, the converse was true. There were more significant intercorrelations for the Asian staff compared to the majority group. This prediction was based on the few studies that show dimensional correlations for different racial groups. It is not clear why there was no similar finding here. Two possible reasons include the fact that these data were from the U.K. while the earlier studies were conducted in the U.S. (the racial dynamics may be different), and that the sample sizes are small. Although the findings do not support the hypothesis, they are nonetheless interesting. There appears to be some halo within the Asian sample (reflected by the number of intercorrelations), which is particularly significant to note since there were no differences in overall ratings between this group and the other two groups.

Research questions 1 and 2 addressed group differences in supervisory ratings while Research Question 3 examined what supervisors *say* about staff in justifying their performance ratings. This addresses Dipboye's (1985) criticism of existing studies, which he contends simply perform group comparisons via analysis of variance, by identifying factors salient to raters at the time of rating. Hypothesis 3a predicted that there would be fewer positive mentions of task factors and more negative mentions of

task factors in supervisors' summaries of minority performance compared to those for majority group members. This hypothesis was not supported. There were significant differences on two of the four positive task factors examined and no differences on the negative task factors. The significant differences were not in the predicted direction, however. Supervisors mentioned positive task factors (sales and execution of work) *more* for Asian staff. This was not the case for black staff. However, when positive and negative mention of task factors were combined, supervisors mentioned sales significantly more for both black and Asian staff. Thus this factor only reached significance for black staff when negative mention was combined with positive mention. Although this hypothesis was not supported in terms of the relative mention of positive versus negative task factors across groups, these results demonstrate the salience of task factors to supervisors in rating minority groups. Task factors are more salient when supervisors are justifying minority performance than when summarizing majority performance.

The lack of support for this hypothesis may be explained by two factors. First, supervisors overwhelmingly made positive comments in justifying their ratings of staff. They gave comparatively less negative feedback. The low frequencies would affect the ability to detect differences in negative comments across groups. Second, the prediction was made based on a review of U.S. research involving predominantly black and white samples as the minority and majority groups. The present study involved two minority groups: Asian and black employees. Initial analyses in the present study suggested differences in results between the two minority groups, resulting in the decision to analyze the groups separately. These groups are also from a different country thus the

interracial dynamics may be different and may not be clearly predictable based on U.S. research. The findings reflect an evolving theme throughout the analyses: supervisors are more likely to make positive comments in justifying their ratings of Asian staff irrespective of the level of overall performance. This may suggest the existence of positive stereotypes regarding Asian staff. As the findings for black staff only reached significance when negative comments were included, this suggests an emphasis on task factors for this group rather than positive stereotyping.

Hypothesis 3b predicted that supervisors would emphasize contextual performance factors more for minority staff than majority staff in their written summaries of performance. This hypothesis received minimal support. There was a significant difference on only one of the 6 factors examined, Adaptability. Supervisors were more likely to describe Asian staff as adaptable, compared to the other two groups, regardless of overall performance rating. It is interesting to note that this is the only factor that supervisors emphasized more for one group relative to another. Adaptive performance has recently been identified as an important dimension of performance quite distinct from task and contextual performance (J. P. Campbell, 1999). Adaptive performance has been described as flexibility, the capacity to cope with change (Hesketh & Neal, 1999) and how individuals self-manage their learning (London & Mone, 1999). There were no significant differences between the majority and black staff on the contextual factors examined.

Essentially, analyses for Research Question 3 show differences in the relative emphases of one aspect of task performance for black versus white staff; and contrary to prediction, no difference in the relative emphasis of contextual performance between

these two groups at the same level of overall performance. For Asian staff, supervisors emphasized two positive task and one contextual factor at the same overall performance level. These findings for task performance suggest positive stereotyping of Asian staff while the findings for contextual performance for the Asian group minimally support predictions.

Research Question 4 sought empirical support for stereotyping and supervisor liking as theoretical explanations for differences in overall performance ratings across groups. It was predicted (Hypothesis 4a) that supervisors would emphasize negative indices of performance in justifying their ratings of minorities; so called because these indicators assess negative behavior rather than positive behavior. Examples are absences, accuracy (in banking terms, shortages and overages), punctuality, and sickness record. This hypothesis received partial support for one minority group. Supervisors were more likely to mention Accuracy in their description of Asian staff's performance. There was no support in the black sample. The other 3 factors examined had low frequencies for positive mentions which might explain the non-significant result, and even lower frequencies for negative mentions. Supervisors were less likely to mention attendance and punctuality in justifying ratings. Support on only one factor for Asian staff and the lack of support for black staff suggests that this finding may reflect positive stereotyping of Asian staff.

Hypothesis 4b predicted differences across groups, at the same level of performance, in supervisors' tendency to make positive comments in their written justifications of performance. This hypothesis was supported. Supervisors were more

likely to make positive comments in general about Asian staff than the other groups at the same level of overall performance.

Hypothesis 4c predicted an emphasis on liking factors by supervisors for majority staff compared to minority staff. This hypothesis was not supported given the direction of the group differences found. Rather than a higher occurrence of liking factors for majority staff, consistent with the clear theme that has emerged, supervisors were more likely to praise Asian staff for their performance and to describe them as 'positive' irrespective of their overall performance rating.

Finally a general research question explored the content of performance factors used by supervisors in justifying performance. These factors were examined across groups at different categories of performance, good, average, and poor. This analysis suggests that sales and helpful behavior were the most salient factors in justifying ratings of good performers across groups. This was particularly true of black staff. Sales was also mentioned relatively more frequently as a negative factor for black staff who performed at the average level. This supports the interpretation of the salience of this factor for black staff. At average performance, Accuracy was the most salient factor for Asians and majority staff while sales and helpful behavior appeared most frequently for black staff. The emphasis on Accuracy for Asian staff was considerably stronger than for white staff suggesting that Accuracy is salient to supervisors in rating this group and may reflect a stereotype on which this group is evaluated.

Conclusions

These data suggest that ethnic minorities generally fare less well than majority staff on overall ratings of performance. In contrast to past research, the focus was on

explication of the performance rating construct rather than mere identification of group effects in ratings. The central research question raised here is whether performance ratings are comparable across groups in meaning or whether they may reflect different underlying constructs. Correlational and regression analyses found that the same performance dimensions do not predict overall performance ratings across ethnic groups. Although the results of this study were not always in the predicted direction, the principle of the central question is upheld as the findings suggest that ratings may indeed have different underlying meanings across groups. Thus, bias, or criterion contamination, can be inferred (Kraiger & Ford, 1990).

The "surprise" finding and consistent theme throughout the analyses was the differences in the relationships among ratings (dimensional, overall, and intercorrelations) for Asian staff compared to their black and majority counterparts; and the differences in factors reported by supervisors in their justification of their ratings for this group. This was also the only group for whom there was no significant relationship between tenure and overall performance. They achieved their ratings and comments irrespective of length of service and experience. In contrast to the other two groups, regression analyses did not identify any predictors of overall performance rating from supervisory dimensional ratings.

One interpretation of the general findings for Asian staff is that supervisors hold positive beliefs about this group, which are reflected in their positive comments and praise at the same level of performance as the other groups. Personal beliefs can be distinguished from cultural stereotypes (Devine, 1989). Cultural stereotypes represent a learned set of associations that link a set of characteristics with a group (Devine & Elliot,

1995). Personal beliefs regarding a group may or may not be congruent with these stereotypes. Research has demonstrated that there is more congruence between personal beliefs and stereotypes in high prejudiced individuals compared to low prejudiced individuals (Devine, 1989). There are no data or available research that identify the cultural stereotype of Asians in Britain; thus it is not possible to determine the extent to which supervisors' comments may reflect stereotypic characteristics. However, there is some research to show that black and Asian women managers in Britain have different experiences in organizations based on the expectations of colleagues and superiors regarding both their gender and ethnicity (Davidson, 1995). These expectations reflect cultural stereotypes. For example, Asian women are expected to be passive while large black women are expected to be aggressive (Davidson, 1995). More data on the nature of stereotypes regarding Asian males and females are needed to determine the extent to which they may be reflected here. This would be a task for future research. It can only be concluded here that supervisors' comments represent their beliefs which may reflect cultural stereotypes.

As there were no significant differences between Asians and the majority group in overall ratings for year 3 (the year for which narrative summaries were available), it is not clear whether these positive beliefs affected overall performance ratings for this group, particularly given that there were significant differences the two years prior. Supervisors' positive beliefs did not serve to increase this group's overall rating relative to the majority group who did not receive similar positive comments. One potential explanation for this finding is a theory suggested by Kraiger and Ford (1990). Positivity bias (Pettigrew, 1979), states that majority group raters inflate ratings of their own group

members because raters consider job-irrelevant factors in rating these individuals. Thus bias may be due to inflated ratings of majority group members rather than the assignment of lower ratings to minorities. The absence of significant differences between Asians and the majority group despite the positive beliefs may be due to positivity bias in the majority group ratings. There were also no significant differences between Asian and black staff's overall ratings indicating a smaller gap between Asian-black ratings than majority-black ratings. In short, supervisors hold positive beliefs about Asian staff which puts their ratings on par with majority staff but not high enough to be different from the other minority group. Future research might examine supervisors' justifications in a sample with significantly different overall ratings to clarify the relationship between raters' beliefs and performance ratings.

In summary, these findings support the argument that the meaning of ratings may not be comparable across groups and may reflect bias.

One limitation of the current study is the small sample sizes for the black and Asian groups. Despite this, there were clear differences across groups on some variables. In the case of the stepwise regression analyses, in particular, cross-validation is important to determine the usefulness of the model obtained for black staff (Myers & Well, 2003; Wilkinson, 1979). Another concern was possible heterogeneity of variance on some dependent variables and violation of the underlying assumptions of F in the ANCOVAs performed. It can be argued, however, that it is legitimate to examine variance as an effect in its own right (Collyer & Enns, 1986). Heterogeneity in variance may result from floor or ceiling effects due to the measuring instrument (Myers & Well, 2003) which in this case is the supervisor. Such effects are of empirical interest in a study such

as this one investigating group differences in evaluation. It is instructive to know for example, that there are ceiling effects for ratings of one group relative to another.

An interesting finding which also represented a limitation is the fact that supervisors gave little negative feedback. If stereotypes of black individuals are negative as suggested by research in the U.S. (Devine & Elliot, 1995) and U.K. research (Davidson, 1995; Jenkins, 1988), the fact that few differences were found in supervisors' comments for the black group may be due to the low frequency of negative comments by supervisors. The preponderance of positive comments allowed positive stereotypes of Asians to emerge. The significant differences in supervisors' evaluations of black staff were reflected in their overall and dimensional ratings rather than in supervisors' comments. This raises two issues: 1) whether and how supervisors provide feedback to those rated as poor performers. It has been suggested that raters who are ambivalent towards minorities may find it difficult to give them negative feedback (Dipboye, 1985). One empirical study has found an inverse relationship between black performance ratings and white raters' propensity to give black ratees feedback compared to white ratees (Feild & Holley, 1977). 2) Under what conditions (if any) negative beliefs are reported.

Implications

The primary contribution of the present study is support for the argument that performance ratings may represent different constructs across groups and as such may reflect criterion contamination. The data clearly show that intercorrelations among ratings differ among groups (at the same level of performance); that different dimensional ratings predict overall performance rating across groups; and that the factors salient to supervisors in justifying their ratings of staff vary across groups. Oppler et al. (1992)

have called for research in which ratings are correlated with a wider variety of measures including positive and negative indicators of performance. This study addresses this call to some extent by providing some data on the interrelationships among different types of supervisory ratings and supervisors' summaries of performance. Oppler et al. also point out that differential information-processing theories such as Dipboye's stereotype-fit remain untested. Although Dipboye's model is not directly tested, the results provide some evidence that different factors (which may include stereotypes) are salient to supervisors when justifying overall ratings for different groups.

There are implications for both research and practice. In terms of research, the finding that overall performance has different predictors across ethnic groups has implications for archival studies where overall performance is computed based on mean dimensional performance. The relationship between dimensional ratings and overall rating is not consistent across groups. This is clearly important to know. Validity coefficients or mean differences determined based on such computations assume the same regression model across groups. These studies should be examining the components of the criterion space rather than aggregating selected elements. These findings also suggest that combination of dimensional ratings across jobs and organizations is likely to obscure meaningful differences across groups; differences that would add to our understanding of the performance construct.

Supervisory ratings are typically the criterion used in validation research. If ratings are not comparable across groups, such bias would call into question the relationship between ratings and predictors used in such research. Schmidt and Hunter's (1998) model of performance ratings primarily uses supervisory ratings as the criterion.

The assumption is that these ratings are true measures of performance. Emphasis is placed on refining the predictors in the model rather than examining the criterion. Schmidt and Hunter's research has been criticized for their failure to consider a broader range of predictors in their model of ratings (Ferris et al., 1994). The findings here raise questions about the precise meaning of supervisory ratings and what the correlations between the predictors and criterion in their model actually reflect. This would also have implications for the use of selection processes based on their model.

Implications for practice include the fact that ethnic minorities in Britain experience direct and indirect discrimination both in terms of entry and upward mobility in organizations (Jenkins, 1988). Like their U.S. counterparts, they are subject to what has been termed the "glass ceiling". Davidson (1995) contends that ethnic minority women (black and Asian) experience double discrimination and are more likely to encounter a "concrete ceiling" (p.35). If supervisors' ratings of performance have different meanings across ethnic groups, this has implications for the validity of ratings made and the subsequent decision-making.

The fact that black staff received significantly lower overall ratings yet supervisors' comments were not different from those made in justifying majority ratings would have implications for performance feedback. Supervisors should be trained to give feedback to poor performers. This may also impact ratings as raters would be forced to think more actively (consciously) about their reason (justifications) for their ratings at the time of rating.

This research also has implications for other minority ethnic groups other than those studied here. In Britain, for example, the Race Relations Act also protects

individuals from discrimination based on sectarianism, nationality, and ethnicity. Irish job applicants, for example, have been asked about drinking habits during interviews (Income Data Services, 1990) and Muslims refused positions based on their attire (Income Data Services, 1996). If raters are using implicit theories that vary across groups, this would have implications for any minority racial, ethnic or religious group. These implicit theories may be positive as they were here which may serve to disadvantage other groups in the organization for which there were negative or even neutral beliefs.

A key strength to the study is the use of real performance data used for personnel decision making rather than laboratory or field research data. The fact that different factors are salient in rating different groups has implications for understanding the rating process in general. For example, it is conceivable that these factors may change as a function of other variables such as job, gender, and organizational context suggesting the use of different implicit theories of performance under different conditions.

Future Research

There are several avenues for future research. At a minimum, this study should be replicated using larger sample sizes to confirm some of the findings here. This is a consistent issue with research investigating group differences. This may be more feasible in the U.S., for example, where there are large percentages of racial minorities. Future studies might examine ratings and supervisory justifications for other groups such as Hispanics and African-Americans versus the majority group in the U.S. to determine whether supervisors' tendency to provide primarily positive feedback is consistent and to identify any differences from the relationships studied here.

Future research should seek to clarify the performance factors identified here and their interrelationships. One clear need is a theoretical understanding of the different predictors of overall performance rating across ethnic groups. Why do supervisors focus on one set of predictors for one ethnic group compared to others?

In this study, comparison of means among the different minority groups suggested the need to analyze these groups separately. As no specific predictions were made regarding differences between the two minority groups, or each minority group vis â vis the majority, analyses at this level represented an exploratory approach. The identification of unexpected and interesting findings, as was the case here, generates planned comparisons for future research (Keppel, 1982). Future research should use a confirmatory approach to compare performance ratings of black and Asian and Asian and majority staff and to compare supervisory comments. For example, will the generally positive comments for Asian staff be replicated? One interesting finding was supervisors' emphasis on Adaptability in describing Asian staff. Confirmatory analyses might examine whether this result is replicated.

Related to this, is the need to clarify differences, if any, in task and contextual elements of performance. Preliminary studies have found inconsistent results. Differences have been found in task performance and some contextual dimensions (Hauenstein et al, 2002). This study found differences in Adaptability, a dimension that has been distinguished from task and contextual performance (J. P. Campbell, 1999). Future studies are needed to clarify any differences as this may have implications for addressing adverse impact in hiring (Hattrup et al, 1998) and would add to our understanding of supervisor's maps of performance for different groups as well as where

any true differences in performance might lie. Factor analytic studies using larger samples would be useful in exploring the factor structure of ratings for different groups.

In the present study, level of performance was determined by the supervisor's overall performance rating. Future research should examine the relationships among dimensional, overall, and "objective" measures of performance, such as actual shortages and overages (as in the case of bank tellers) to identify any differences among these relationships across groups. This would also allow examination of supervisors' justifications at different levels of objective performance.

Future research should also explore the generalizability of these findings to other jobs and organizations. One of the goals of this study was to examine real performance ratings in one organization as opposed to across organizations as in the large archival studies that are typical in this literature. Large archival datasets may obscure meaningful differences in raters' rating strategies and the nature of the performance construct across groups. Future research should extend this study by examining ratings and summaries for other jobs and other organizations to determine whether the findings here are replicated in other jobs and organizations. For example, are these relationships among ratings and views of Asian staff found here consistent across jobs and organizations? Stereotype-fit for example, would predict that raters compare stereotypes of effective performers with stereotypes of the relevant group to determine the individual's fit with the job (Dipboye, 1985). Would different stereotypes or personal beliefs be salient as a function of the job vis â vis the ratee group? This is an empirical question.

Finally, there was little information about the raters in this sample. For example, rater race was unknown although they primarily white. Future research should

investigate rater-ratee race interactions. Specifically, is the underlying meaning of performance ratings different for different rater-ratee race combinations? This would provide insight to whether raters' implicit theories of performance vary as a function of both rater and ratee race; and whether there are more differences within or across racial groups in implicit theories of performance.

A Final Note

Cronbach and Meehl (1955) sum up the central issue here in their discussion of the validity of criteria. They provide the following example:

> In some situations the criterion is no more valid than the test. Suppose, for example, that we want to know if counting the dots on Bender-Gestalt figure five indicates "compulsive rigidity," and take psychiatric ratings on this trait as a criterion. Even a conventional report on the resulting correlation will say something about the extent and intensity of the psychiatrist's contacts and should describe his qualifications (e.g., diplomate status? analyzed?).
>
> Why report these facts? Because data are needed to indicate whether the criterion is any good….In [a] study where a construct is the central concern, any distinction between the merit of the test and criterion variables would be justified only if had already been shown that the psychiatrist's theory and operations were excellent measures of the attribute. (pp.284-285)

Why do we not set the same standard for performance ratings? In research, we combine ratings across jobs, purposes, and organizations with no examination of the

validity of the source of the rating or the appropriateness of the rater. Why do we take for granted validity of the criterion? The results of this research suggest we cannot afford such complacency.

REFERENCES

American Psychological Association. (1991). In the supreme court of the United States: Price Waterhouse v. Ann B. Hopkins: (Amicus curiae brief for the American Psychological Association). *American Psychologist, 46*, 1061-1070.

American Psychological Association. (2003). *In the supreme court of the United States: Grutter v. Bollinger, et al. and Gratz and Hamacher v. Bollinger, et al. (Brief amicus curiae of the American Psychological Association in support of respondents)*. Retrieved March, 2003

Austin, J. T., & Villanova, P. (1992). The criterion problem. *Journal of Applied Psychology, 77*(6), 836-874.

Baehr, M. E., Saunders, D. R., Froemel, E. C., & Furcon, J. E. (1971). The prediction of performance for black and for white police patrolmen. *Professional Psychology, 2*, 46-57.

Barnes-Farrell, J. L. (2001). Performance appraisal: Person perception processes and challenges. In *How People Evaluate Others in Organizations*. (pp. 135-153). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology, 57*(2), 101-109.

Beatty, R. W. (1973). Blacks as supervisors: A study of training, job performance, and employers' expectations. *Academy of Management Journal, 16*(2), 196-207.

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61*(1), 80-84.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Borman, W. C. (1987). Personal constructs, performance schemata, and 'folk theories' of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behaviour and Human Decision Processes,, 40*, 307-322.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: the meaning for personnel selection research. *Human Performance, 10*(2), 99-109.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*(6), 863-872.

Boyce, A. M., Pratt, A., Bauer, C. C., Amelio, S. L., & Baltes, B. B. (2002). *Stereotypes of black male managers and professors: Scale development.* Paper presented at the Society for Industrial and Organizational Psychology, Toronto, Canada.

Brugnoli, G. A., Campion, J. E., & Basen, J. A. (1979). Racial bias in the use of work samples for personnel selection. *Journal of Applied Psychology, 64*(2), 119-123.

Campbell, J. P. (1999). The definition and measurement of performance in the new age. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 258-299). San Francisco: Jossey-Bass.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 56 (2), 81-89.

Campbell, J. P., Gasser, M. B., & Oswald, F. (1996). The substantive nature of performance variability. In K. R. Murphy (Ed.), *Individual Differences and Behavior in Organizations.* (pp. 258-299). San Francisco, CA: Jossey-Bass.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey Bass.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six-year study.* (No. PR-73-27). Princeton: Educational Testing Service.

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology, 71*, 672-678.

Cardy, R. L., & Dobbins, G. H. (1994). Performance appraisal: The influence of liking on cognition. In C. Stubbart, J. R. Meindel & J. F. Porac (Eds.), *Advances in Managerial Cognition and Organisational Information Processing*. London: JAI Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, New Jersey: Erlbaum

Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. *Human Resources Management Review, 10*, 25-44.

Collyer, C. E., & Enns, J. T. (1986). *Analysis of Variance: The Basic Designs*. Chicago: Nelson-Hall.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52 (4), 281-300.

Davidson, M. J. (1995). Living in a bicultural world - the role conflicts facing the black ethnic minority woman manager. *International Review of Women and Leadership*(1), 22-36.

Deaux, K. (1976). Sex: A perspective on the attribution process. In J. H. Harvey, W. J. Ickes & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1). Hillsdale, N.J.: Erlbaum.

Deaux, K., & Emswiller, T. (1974). Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. *Journal of Personality and Social Psychology, 29*, 80-85.

DeJung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology, 46*(5), 370-374.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance., 33*, 360-396.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality & Social Psychology, 56*(1), 5-18.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality & Social Psychology Bulletin, 21*(11), 1139-1150.

Devine, P. G., & Elliot, A. J. (2000). Are racial stereotypes really fading? The Princeton trilogy revisited. In *Stereotypes and prejudice: Essential readings.* (pp. 86-99). Philadelphia, PA, US: Psychology Press.

Dipboye, R. L. (1985). Some neglected variables in research on discrmination in appraisals. *Academy of Management Review, 10*, 116-127.

Dobbins, G. H., & Russell, J. (1986). The biasing effects of subordinate likeableness on leaders' attributions and corrective actions. *Personnel Psychology, 39*, 759-777.

DuBois, 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*(2), 205-211.

Erdener, C. B., & Dunn, C. P. (1990). Content analysis. In A. S. Huff (Ed.), *Mapping Strategic Thought* (pp. 291-300). Chichester: John Wiley and Sons.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology, 24*, 609-636.

Feild, H. S., Bayley, G. A., & Bayley, S. M. (1977). Employment test validation for minority and nonminority production workers. *Personnel Psychology, 30*, 37-46.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.

Ferris, G. R., Judge, T. A., Rowland, K. M., & Fitzgibbons, D. E. (1994). Subordinate influence and the performance evaluation process: Test of a model. *Organizational Behavior and Human Decision Processes, 58*, 101-135.

Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology, 44*, 155-194.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In *The handbook of social psychology, Vol. 2 (4th ed.).* (pp. 357-411). New York, NY, US: McGraw-Hill, New York.

Fiske, S. T., Bersoff, D. N., Borgida, E., Deaux, K., & al, e. (1991). Social science research on trial: Use of sex stereotyping research in Price Waterhouse v. Hopkins. *American Psychologist, 46*(10), 1049-1060.

Ford, J. K., Schechtman, S. L., & Kraiger, K. (1986). Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. *Psychological Bulletin, 99*(3), 330-337.

Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology, 27*, 207-223.

Greenhaus, J. H., & Gavin, J. F. (1972). The relationship between expectancies and job behavior for white and black employees. *Personnel Psychology, 25*(3), 449-455.

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organisational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal., 33*(1), 64-86.

Guion, R. M. (1983). Comments on Hunter. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance Measurement and Theory* (pp. 267-275). Hillsdale, N.J.: Erlbaum.

Hall, F. S., & Hall, D. T. (1976). Effects of job incombents' sex and race on evaluations of managerial performance. *Academy of Management Journal, 19*, 476-481.

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*(6), 705-711.

Hattrup, K., O'Connell, M. S., & Wingate, P. H. (1998). Prediction of multidimensional criteria: distinguishing task and contextual performance. *Human Performance, 11*(4), 305-319.

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82*, 656-664.

Hauenstein, N. M. A., Sinclair, A. L., Robson, V., Quintella, Y., & Donovan, J. J. (2002). Ratee race effects in performance: Task performance versus contextual performance. On *SIOP Conference*. Toronto, Canada: SIOP.

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

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811-826.

Hesketh, B., & Neal, A. (1999). Technology and performance. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 21-55). San Franciscso: Jossey-Bass.

Hobson, C. J., & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review., 8*(4), 640-649.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology, 29*, 13-30.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance and supervisor ratings. In F. J. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance Measurement and Theory* (pp. 257-266). Hillsdale, N.J.: Erlbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340-362.

Ilgen, D. R. (1983). Gender issues in performance appraisal: A discussion of O'Leary and Hansen. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Performance Measurement and Theory*. Hillsdale, N.J.: Erlbaum.

Ilgen, D. R., & Youtz, M. A. (1990). Factors affecting the evaluation and development of minorities in organisations. In G. R. Ferris & K. M. Rowland (Eds.), *Performance Evaluation, Goal Setting, and Feedback.* London: JAI Press.

Income Data Services (1990). *Race Discrimination*. Old Woking, Surrey: Income Data Services.

Income Services (1996). *Race Discrimination* (No. Series 2, No. 9). Old Woking, Surrey: Income Data Services.

Jeanquart-Barone, S. (1996). Implications of racial diversity in the supervisor-subordinate relationship. *Journal of Applied Social Psychology, 26*(11), 935-944.

Jenkins, R. (1988). Discrimination and equal opportunity in employment: Ethnicity and "race" in the United Kingdom. In D. Gallie (Ed.), *Employment in Britain*. Oxford: Basil Blackwell.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervsior judgments of overall performance. *Journal of Applied Psychology, 86*(5), 984-996.

Jones, T. (1993). *Britain's Ethnic Minorities: An Analysis of the Labour Force Survey*. London: Policy Studies Institute.

Kanter, R. M. (1977). *Men and women of the corporation*. New York: Basic Books.

Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook* (Second ed.). Englewood Cliffs: Prentice-Hall, Inc.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R., & Katzell, R. A. (1968). *Testing and Fair Employment*. New Yok: New York University Press.

Klimoski, R. J., & Donahue, L. M. (2001). Person perception in organizations: An overview of the field. In *How People Evaluate Others in Organizations*. (pp. 5-43). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal Of Applied Psychology, 70*(1), 56-65.

Kraiger, K., & Ford, J. K. (1990). The relation of job knowledge, job performance, and supervisory ratings as a function of ratee race. *Human Performance, 3*, 269-279.

Lance, C. E. (1994). Test of a latent structure of performance ratings derived from Wherry's (1952) theory of ratings. *Journal of Management, 20*, 757-771.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Landy, F. J., & Farr, J. L. (1983). *The Measurement of Work Performance*. New York: Academic Press.

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational and Organizational Psychology, 73*, 67-85.

Lefkowitz, J., & Batista, M. (1995). Potential sources of criterion bias in supervisor ratings used for test validation. *Journal of Business and Psychology, 9*(4), 389-414.

Lefkowitz, J., & Battista, M. (1995). Potential sources of criterion bias in supervisor ratings used for test validation. *Journal Of Business and Psychology, 9*(4), 389-414.

LePine, J. A., Hanson, M. A., Borman, W. C., & Motowidlo, S. J. (2000). Contextual performance and teamwork: Implications for staffing. In *Research in Personnel and Human Resources Management* (Vol. 19, pp. 53-90): Elsevier Science, Inc.

London, M., & Mone, E. (1999). Continuous learning. In D. R. Ilgen & E. D. Pulakos (Eds.), *The Changing Nature of Performance:Implications for Staffing, Motivation and Development* (pp. 119-153). San Francisco: Jossey-Bass.

Lord, R. G., & Maher, K. J. (1989). Cognitive processes in industrial and organizational psychology. In C. L. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organisational Psychology.* New York: John Wiley & Sons Ltd.

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*(3), 598-606.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*(2), 71-83.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79*, 475-480.

Mount, M. K., Hazucha, J. F., Holt, K. E., & Sytsma, M. (1995). *Rater-ratee race effects in performance ratings of managers.* Paper presented at the Academy of Management Annual Meetings, Vancouver, British Columbia, Canada.

Mount, M. K., & Scullen, S. E. (2001). Multisource feedback ratings: What do they actually measure? In M. London (Ed.), *How People Evaluate Others in Organizations* (pp. 155-176). Mahwah: Lawrence Erlbaum Associates.

Mount, M. K., Sytsman, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater-ratee race effects in developmental performance ratings of managers. *Personnel Psychology, 50*, 51-69.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selectin tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823-853.

Myers, J. L., & Well, A. D. (2003). *Research Design and Statistical Analysis* (Second ed.). New Jersey: Lawrence Erlbaum Associates.

Nagle, B. F. (1953). Criterion development. *Personnel Psychology, 6*, 271-288.

Nunnally, J. C. (1978). *Psychometric Theory.* New York: McGraw-Hill.

O'Leary, V. E., & Hansen, R. D. (1983). Performance evaluation: A social-psychological perspective. In F. Landy, S. Zedeck & J. Cleveland (Eds.), *Perormance Measurement and Theory.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Operario, D., & Fiske, S. T. (2001). Causes and consequences of stereotypes in organizations. In *How people evaluate others in organizations.* (pp. 45-62). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology, 77*(2), 201-217.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*(2), 85-97.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*. New York: Holt, Rinehart and Winston.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality & Social Psychology Bulletin, 5*, 461-476.

Powell, G. N., & Butterfield, D. A. (1997). Effect of race on promotions to top management in a federal department. *Academy Of Management Journal, 40*(1), 112-128.

Powell, G. N., & Butterfield, D. A. (2002). Exploring the influence of decision makers' race and gender on actual promotions to top management. *Personnel Psychology, 55*(2), 397-428.

Prewett-Livingston, A. J., Feild, H. S., Veres, J. G., & Lewis, P. M. (1996). Effects of race on interview ratings in a situational panel interview. *Journal of Applied Psychology, 81*(2), 178-186.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance., 9*(3), 241-258.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of ratee race, ratee gender, and rater level effects. *Human Performance, 92*(2), 103-119.

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*(5), 770-780.

Robbins, T. L., & DeNisi, A. S. (1993). Moderators of sex bias in the performance appraisal process: A cognitive analysis. *Journal of Management, 19*, 113-126.

Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology, 84*(5), 815-822.

Ryan, G. W., & Bernard, H. R. (2000). Data Management and Analysis Methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (Second ed., pp. 769-802). Thousand Oaks: Sage.

Sackett, P. R., & DuBois, C. L. Z. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76*(6), 873-877.

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *The Journal of Applied Psychology, 57*, 95-100.

Schein, V. E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *The Journal of Applied Psychology, 60*, 340-344.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1/2), 187-210.

Schmidt, F. L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research. *American Psychologist, 36*, 1128-1137.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample, performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432-439.

Schmidt, F. L., & Johnson, R. H. (1973). Effect Of Race On Peer Ratings In An Industrial Situation. *Journal of Applied Psychology, 57*(3), 237.

Schmitt, N., & Hill, T. E. (1977). Sex and Race Composition of Assessment Center Groups as a Determinant of Peer and Assessor Ratings. *Journal of Applied Psychology, 62*(3), 261-.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65*(4), 428-435.

Schmitt, N., & Noe, R. (1986). Personnel selection and equal employment opportunity. In C. L. Cooper & I. Robertson (Eds.), *International Review of Industrial and Organizational Psychology.* (pp. 71-115). New York: Wiley.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Stewart, D. L., & Perlow, R. (2001). Applicant race, job status, and racial attitude as predictors of employment discrimination. *Journal of Business and Psychology, 16*(2), 259-275.

Tajfel, H., & Forgas, J. P. (2000). Social categorization: Cognitions, values and groups. In C. Stangor (Ed.), *Stereotypes and Prejudice: Essential Readings* (pp. 49-63). Philadelphia: Psychology Press.

Thompson, D. E., & Thompson, T. A. (1985). Task-based performance-appraisal for blue-collar jobs - evaluation of race and sex effects. *Journal Of Applied Psychology, 70*(4), 747-753.

Tomkiewicz, J., Brenner, O. C., & Adeyemi-Bello, T. (1998). The impact of perceptions and stereotypes on the managerial mobility of African Americans. *The Journal of Social Psychology, 138*(1), 88-92.

Toole, D. L., Gavin, J. F., Murdy, L. B. & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology*, *55,* 661-672.

Waldman, D. A., & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76*(6), 897-901.

Wendelken, D. J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? *Journal of Applied Psychology, 66*(2), 149-158.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86(1)*, 168-174.

Wilson, K. Y. (1995). Appraisal: A fair assessment?, *CRE Connections*. London: Commission for Racial Equality.

Wollowick, H. B., Greenwood, J. M., & McNamara, W. J. (1969). *Psychological testing with a minority group population.* Paper presented at the APA.

APPENDIX A

PERFORMANCE APPRAISAL FORM

Name: Job Title:

PART I

Part 1 of the review considers performance over the review period

SUMMARY OF PERFORMANCE

PERFORMANCE RATING - Overall performance during the last review period

O= Outstanding	H= High Achievement	G= Good	I= Improvement Required	U= Unacceptable	R= Unrated	P= Progress Review

SKILLS – Note the level achieved in relation to the level required.

	Level Achieved	Required		Level Achieved	Required
Business Awareness	0	0	Initiative	0	0
Business Development	0	0	Numeracy	0	0
Communications - Verbal	0	0	Planning & Org	0	0
Communications - Written	0	0	Team Work & Lead	0	0
Customer Service	0	0	Use of Systems	0	0
Decision Making	0	0	Work Standards	0	0

Period of Review: Review Date:

Current Branch: Grade:

No of days of sickness since last annual review:

KNOWLEDGE (PEFORMANCE REVIEW ONLY) Comment on achievement in the main areas of knowledge, eg products/procedures/systems

STRENGTHS & AREAS FOR DEVELOPMENT
Areas of Strength (PERFORMANCE REVIEW ONLY)
Areas for Development (PERFORMANCE REVIEW ONLY)

PART 2
ACTION PLANNING FOR THE NEXT REVIEW PERIOD
Part 2 of the review considers what needs to be achieved over the next review period

PERFORMANCE IMPROVEMENT/DEVELOPMENT PLAN (Optional where performance is at required standard or above (G/H/O), mandatory where it is below (U/I)
PERFORMANCE IMPROVEMENT/DEVELOPMENT AREA
1 2 3 4 5

6
ACTION TO BE TAKEN 1 2 3 4 5 6

BY WHEN	
1 2 3	4 5 6

WHO RESPONSIBLE	
1 2 3	4 5 6

SIGNATURES	
Manager/Supervisor Signature: Name:	Date:
Individual's Comments: Signature:	Date: Name:
Countersignature: (PERFORMANCE REVIEW ONLY) Name:	Date:

APPENDIX B

APPRAISAL FORM – EXAMPLE SUMMARY OF PERFORMANCE

SUMMARY OF PERFORMANCE

"Joan" has successfully achieved the core standards of behavior; provides a polite and personalized cashiering service (use of names or sir/madam) and is extremely helpful to colleagues especially the Counter Manager (pay in machine/Fastbank/night safes). Her customer interaction has achieved MPFS commission to Nov 1440 – Well Done. Joan provided "Main Street" branch with cashiering cover in Sept. Joan continues to supervise the running of our local Midbank, undertaking regular visits and personally overseeing the input of transactions and account opening, often working late to ensure her job is up straight. Two fraud certificates have been received, but during Sept – Nov Joan's cashiering differences had reached an unacceptable level totaling $536.93 short. However, this does appear to have been an isolated period caused by extra pressure from both inside and outside work.

APPENDIX C

PARTIAL CORRELATIONS AMONG DIMENSION RATINGS

Dimen.	2	3	4	5	6	7	8	9	10	11	12
1.BusA	44**	35**	15*	32**	40**	42**	17**	24**	31**	12	31**
2.BusD		29**	13	21**	29**	32**	30**	23**	07	18**	05
3.CV			28**	53**	46**	45**	17**	39**	55**	11	43**
4.CW				10	32**	21**	35**	23**	26**	25**	28**
5.CstSv					40**	40**	00	25**	44**	04	48*
6.DM						60**	30**	45**	38**	06	46**
7.Init							22**	40**	37**	01	43**
8.Num								22**	09	29**	07
9.Pl&O									44**	00	48**
10.Team										05	62**
11.Usys											08
12.Stnd											

Note. n=203-218. *p < .05 . **p < .00.

Table 38. Partial Correlations Among Dimension Ratings with Tenure and Overall Rating Partialled Out – Full White Sample

APPENDIX D

ANALYSIS OF COVARIANCE TABLES

Source	*df*	F	p
Overall Rating	1	17.21	.00
Ethnic Group	2	3.44	.03
Error	663	(.153)	

Note. Values in parentheses represent mean square errors.

Table 39. Analysis of Covariance for Comparison of Positive Mention of Sales by Group

Source	*df*	F	p
Overall Rating	1	8.08	.00
Ethnic Group	2	.378	.68
Error	663	(6.883E-02)	

Note. Values in parentheses represent mean square errors.

Table 40. Analysis of Covariance for Comparison of Positive Mention of Productivity Cluster by Group

Source	*df*	F	p
Overall Rating	1	2.50	.11
Ethnic Group	2	2.79	.06
Error	663	(.294)	

Note. Values in parentheses represent mean square errors.

Table 41. Analysis of Covariance for Comparison of Positive Mention of Knowledge and Learning New Tasks Cluster by Group

Source	*df*	F	p
Overall Rating	1	1.18	.17
Ethnic Group	2	4.66	.06
Error	663	(.722)	

Note. Values in parentheses represent mean square errors.

Table 42. Analysis of Covariance for Comparison of Positive Mention of Execution of Work Cluster by Group

Source	*df*	F	p
Overall Rating	1	6.64	.01
Ethnic Group	2	.914	.40
Error	663	(.061)	

Note. Values in parentheses represent mean square errors.

Table 43. Analysis of Covariance for Comparison of Negative Mention of Sales by Group

Source	*df*	F	p
Overall Rating	1	/02	.88
Ethnic Group	2	1.36	.25
Error	663	(.012)	

Note. Values in parentheses represent mean square errors.

Table 44. Analysis of Covariance for Comparison of Negative Mention of Knowledge and Learning New Tasks Cluster by Group

Source	*df*	F	p
Overall Rating	1	17.90	.00
Ethnic Group	2	.539	.58
Error	663	(.095)	

Note. Values in parentheses represent mean square errors.

Table 45. Analysis of Covariance for Comparison of Negative Mention of Execution of Work by Group

Source	*df*	F	p
Overall Rating	1	4.84	.02
Ethnic Group	2	3.29	.03
Error	663	(.200)	

Note. Values in parentheses represent mean square errors.

Table 46. Analysis of Covariance for Comparison of Positive and Negative Mention of Sales by Group

Source	*df*	F	p
Overall Rating	1	8.08	.00
Ethnic Group	2	.378	.68
Error	663	(.069)	

Note. Values in parentheses represent mean square errors.

Table 47. Analysis of Covariance for Comparison of Positive and Negative Mention of Productivity by Group

Source	*df*	F	p
Overall Rating	1	2.25	.13
Ethnic Group	2	2.49	.08
Error	663	(.315)	

Note. Values in parentheses represent mean square errors.

Table 48. Analysis of Covariance for Comparison of Positive and Negative Mention of Knowledge and Learning New Tasks by Group

Source	*df*	F	p
Overall Rating	1	.032	.85
Ethnic Group	2	4.95	.00
Error	663	(.816)	

Note. Values in parentheses represent mean square errors.

Table 49. Analysis of Covariance for Comparison of Positive and Negative Mention of Execution of Work by Group

	F	df	p
Positive Mention			
Sales, cross sales, referrals	4.05	2	.00
Productivity	1.85	2	.15
Knowledge & Learning New Tasks	8.72	2	.00
Execution of Work	5.13	2	.00
Negative Mention			
Sales, cross sales, referrals	4.25	2	.01
Productivity	0	0	0
Knowledge & Learning New Tasks	5.53	2	.00
Execution of Work	2.74	2	.06
Positive and Negative Combined			
Sales, cross sales, referrals	9.70	2	.00
Productivity	1.85	2	.15
Knowledge & Learning New Tasks	6.96	2	.00
Execution of Work	3.75	2	.02

Table 50. Levene's test of Equality of Error Variances for Asian, Black, and Majority Groups on Task Performance Factors