REFERENCE REPORTS:

A META-ANALYTIC REVIEW OF PREDICTIVE VALIDITY AND AN EXPERIMENTAL STUDY OF RATING ACCURACY

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Psychology at Massey University

KARL BRANDON PAJO

1996

Massey University Library
Thesis Copyright Form

Title of Thesis:

(1) (a) I give permission for my thesis to be made available to readers in Massey University Library under conditions determined by the Librarian.

~~(b) I do not wish my thesis to be made available to readers without my written consent for ... months.~~

(2) (a) I agree that my thesis, or a copy, may be sent to another institution under conditions determined by the Librarian.

~~(b) I do not wish my thesis, or a copy, to be sent to another institution without my written consent for ... months.~~

(3) (a) I agree that my thesis may be copied for Library use.

~~(b) I do not wish my thesis to be copied for Library use for ... months.~~

Signed 

Date 31/5/96

The copyright of this thesis belongs to the author. Readers must sign their name in the space below to show that they recognise this. They are asked to add their permanent address.

NAME AND ADDRESS DATE

Abstract

Reference reports are a commonly used selection method in New Zealand and overseas. Although popular with practitioners, they have attracted little attention from researchers. To ascertain the predictive validity of reference reports a meta-analytic review was conducted. Results of the preliminary analysis provided a bare-bones estimated mean validity of 0.15, and a fully corrected estimated mean validity of 0.32. Substantial variance remained unaccounted for following corrections for sampling error. Subsequent moderated meta-analyses, based on degree of structure and psychometric soundness of the reference reports, was found to account for much of the variation in observed validity coefficients. Highly structured reports were found to be consistently superior to unstructured reports. Improvements in the validity of highly structured reports can be attributed to the control of leniency in ratings. However, no studies to date have evaluated the accuracy of referees' ratings. Drawing on the performance rating literature, an experimental study examining the influence of scale format, ratee characteristics, rating purpose, and rater affect on the accuracy of ratings was implemented. Asymmetrical, positively toned scales were found to reduce leniency in ratings compared to unstructured and Likert-type rating forms. Raters who expressed liking for the ratee were more lenient in their ratings compared to raters who expressed neutral or antagonistic feelings toward the ratee. No significant effects for rating purpose and ratee characteristics were apparent. Overall, the investigation points to deficiencies in the way reference reports are presently employed, and highlights the need for a more rigorous approach in their development and application.

Acknowledgments

I am indebted to my supervisors, John Podd and Mike Smith, for the support and assistance they offered during the research and writing of this dissertation. The common sense and sound advice they have provided has been much appreciated. I would like to express my special thanks to John, who willingly took on the major supervisory duties when Mike left Massey University to take up another position.

I am very grateful for the encouragement and aid forthcoming from my colleagues in the Departments of Psychology, and Human Resource Management. Not only are they a great bunch of people to work with, but they have proved to be good friends as well. My thanks also go to John Spicer who helped clarify some of the statistical issues that arose throughout the course of the research.

I would also like to express my gratitude to my friends and my family who have been a supportive and understanding crew, especially on those occasions when I couldn't "come out to play."

Finally, it is with deep appreciation that I wish to acknowledge the contribution of my partner, Jacqui. I am very grateful for her constant encouragement, patience, and support during the long nights and weekends when she would have much preferred to see me at home. Thank you.

TABLE OF CONTENTS

Page

Abstract ..... *ii*

Acknowledgements ..... *iii*

Chapter 1

Overview 1

Chapter 2

Literature Review: Reference Reports 5

Surveys of Use 6

Content of Reference Reports 11

Psychometric Issues 15

Reliability 15

Validity 20

Source of the Recommendation 20

Target Population 22

Format and Content 23

Miscellaneous Threats to Validity 27

Meta-Analytic Studies 31

The Meta-Analysis 32

Chapter 3

Meta-Analysis 34

Method 34

Literature Search 34

Decision Rules for Coding Studies into the Meta-Analysis 35

Page

Meta-Analysis Method Used 39
Results 43
Post hoc Analyses 45
Discussion 51

Chapter 4

Literature Review: Performance Rating 60
Rater Training 60
Cognitive Processes in Performance Rating 64
Rating Scale Format 65
Summary 69
Rater Acquaintance and Affect 71
Summary 79
Ratee Characteristics 79
Race 80
Sex 82
Summary 84
Purpose of Rating 86
Rater Motivation 92
Summary 95
Rater Accuracy 96

Chapter 5

Experimental Study 99

Chapter 6

Method 103
Procedural Overview 103
Participants 103
Procedure 104

Page

Experimental Methods 105
Vignette 105
Likability Ratings 108
Demographic Questionnaire 108
Rating Forms 109
Leniency Scale 110
True Scores 112
Dependent Variables 114
Accuracy Measures 114
Error Measures 116
Data Analysis 117
Rescaling of Ratings and Measures 117
Analyses 119

Chapter 7

Results 121

Demographic Questionnaire 121
Tests of Hypotheses 127
Accuracy Measures 131
Error Measures 134
Comparisons Between Measures 136
Regression Analysis of Error Scores 138
Prediction of Leniency 140

Chapter 8

Discussion 147

The Rating Instrument 148
Rater Affect 151
Rating Purpose 157
Ratee Characteristics 161

Page

Criterion Measures 163
Survey of Reference Reports 165
Limitations of the Current Study 166
Future Research 170
Summary and Conclusions 175

References 179

APPENDIX 1

Letter Requesting Participation in the Study 205

APPENDIX 2

Teaching Vignette and Likability Scale 208

APPENDIX 3

Demographic Questionnaire 217

APPENDIX 4

Unstructured Rating Forms Developed for the Referee's Report and for the Performance Appraisal 219

APPENDIX 5

Likert-Type Rating Forms Developed for the Referee's Report and for the Performance Appraisal 222

APPENDIX 6

Asymmetrical Rating Forms Developed for the Referee's Report and for the Performance Appraisal 227

Page

APPENDIX 7

Gender-Typed versions of the Leniency Scale 230

APPENDIX 8

Comparison of Rating Form Accuracy Using the Original Rating Metric 233

LIST OF TABLES

Page

TABLE 1

Studies, sample type, total number of participants and validity coefficients contributed by each study to the meta-analysis 35

TABLE 2

Meta-analysis of the predictive validity of reference reports 44

TABLE 3

Meta-analysis of the predictive validity of reference reports moderated by structure 48

TABLE 4

Meta-analysis of the predictive validity of reference reports moderated by structure and criterion type..... 49

TABLE 5

Mean reliability (average correlations) of judges' ratings for responses on the unstructured rating form..... 117

TABLE 6

Summary of responses to the yes/no items in the demographic questionnaire..... 122

TABLE 7

Descriptive statistics for questions from the demographic questionnaire..... 122

TABLE 8

Significant main effects for type of form from the overall ANOVAs calculated for each teaching dimension..... 128

Page

TABLE 9

Means and standard deviations of ratings of teaching dimensions from three different rating forms 129

TABLE 10

Results from t-test comparing mean ratings from three different forms 130

TABLE 11

Mean accuracy values as a function of purpose and rating form 131

TABLE 12

Results of t-tests comparing mean accuracy values for ratings from three different forms 132

TABLE 13

Mean error values for ratings from three different forms used for performance appraisal and reference report purposes 135

TABLE 14

Correlations among rater error measures 136

TABLE 15

Correlations among rater accuracy measures 137

TABLE 16

Correlations between accuracy and error measures 138

Page

TABLE 17

Results from a standard multiple regression analysis using four error measures to predict each of four accuracy scores 139

TABLE 18

Correlations among variables included in the standard regression analysis for the prediction of leniency in ratings 141

TABLE 19

Results from a standard multiple regression using six personal and contextual variables to predict leniency in ratings..... 143

TABLE 20

Results from the hierarchical multiple regression analysis for the prediction of leniency in ratings after controlling for the effects of rating form (Analysis 1) and for the effects of rating form and Likability (Analysis 2)..... 145

TABLE 21

Referees' access to information about task and interpersonal behaviours and results..... 174

TABLE 22

Results of t-tests comparing mean accuracy values for ratings from three different forms using the original rating metric..... 234

LIST OF FIGURES

Page

FIGURE 1

Distribution of validity coefficients from studies in the meta-analysis 39

FIGURE 2

Rated usefulness of reference reports 124

Chapter 1 Overview

All selection procedures are characterised by a focus on the applicant. Essentially, they entail efforts on the part of employers to elicit information about prospective employees so that they can make informed decisions and choose the "best" individual for the job. Naturally enough, most of these procedures require the direct participation of the applicant. He or she is usually regarded as the primary source of relevant data. Reference reports are unusual in that they do not seek information directly from the prospective employee, but instead rely on judgements and information provided by individuals familiar with the applicant. The function of reference reports appears to be twofold. Factual material provided by third parties can be used to verify information obtained from the applicant. In a sense, reference checks can act as a type of "lie detector" to help determine the "truthfulness" of the applicant. In addition, evaluative judgements obtained from referees can assist in the assessment of the qualifications and capabilities of the candidate. Evaluative data of this type can then be used to identify and make predictions about applicants who are likely to succeed on the job.

Reference reports have received scant attention in the selection literature. The majority of researchers view them as poor predictors with little reliability or validity. However, despite their poor track record, reference reports remain popular with practitioners. Surveys have repeatedly shown that they are one of the most commonly used selection methods. The widespread use of reference reports makes it all the more alarming that they have been largely

ignored by researchers. The purpose of the present investigation is to assess the predictive validity of reference reports and to evaluate the rating accuracy of referees.

Chapter 2 begins by reviewing the existing research literature on reference reports. Surveys conducted in New Zealand and overseas are examined for evidence of any trends associated with the use of reference reports in selection. Studies that have addressed psychometric issues, namely reliability and validity, are discussed and critiqued. Leniency in ratings from referees is identified as a commonly cited problem, and the research evidence that is reviewed suggests that reference reports are poor predictors of future job performance. Several potential moderators of reference report validity are discussed, but no clear conclusions regarding their impact is possible. The necessity for a quantitative review is established, and a meta-analytic study proposed. The scarcity of research, and in particular, contemporary research, is highlighted throughout.

Chapter 3 begins by detailing the method used for the meta-analytic review of reference reports. The search procedure and decision rules for including validity coefficients are described, and the meta-analytic formulas proposed by Raju, Burke, Normand, and Langlois (1991) are presented. The chapter then goes on to report the results from the meta-analytic review. An overall analysis of the predictive validity of reference reports is conducted, and is then followed up with post hoc analyses assessing the influence of several moderator variables. Chapter 3 concludes with a discussion of the results

from the meta-analysis. Some limitations are described, and additional factors that could influence the validity of reference reports are identified. Links between reference reports and performance rating research are introduced. It is noted that no studies have directly examined the accuracy of ratings provided by referees.

Unlike reference reports, there has been a considerable amount of research on performance rating. This literature is reviewed in Chapter 4. Particular attention is paid to certain social, situational, and contextual aspects of the rating situation that have been shown to influence rating outcomes for performance appraisals, and that could generalise to ratings by referees. More specifically, research documenting the effects of rater affect, the purpose of rating, and ratee sex and race is described. Results from the meta-analysis suggested that the format and structure of reference reports might influence ratings. Evidence from contemporary research on performance appraisal rating instruments is also reviewed. The potential for positively toned, asymmetrical rating scales to reduce leniency in ratings is highlighted, and the necessity for further research on the rating accuracy of referees is stressed.

Chapter 5 continues on from the review of the performance rating literature to outline the objectives of an experimental study into the rating accuracy of referees. Specific hypotheses are formulated and stated.

The experimental method used to test the hypotheses is presented in Chapter 6. The procedure, participants, materials, dependent measures, and analytical strategy are described.

Results from the experimental study investigating rating accuracy in reference reports and performance appraisals are reported in Chapter 7. Factors contributing to the prediction of leniency in ratings are identified and evaluated using multiple regression. Survey data collected during the experiment are presented and analysed.

Chapter 8 presents a detailed discussion of the experimental results and places them in the wider context of research on reference reports and rating accuracy. Limitations of the current investigation are also discussed along with suggestions for future research. A summary of the overall research programme and the conclusions that can be drawn bring the thesis to a close.

Chapter 2

Literature Review: Reference Reports

A quick perusal of the literature dealing with personnel selection reveals that reference reports are known by a variety of different names. The most common of these include reference check, letter of recommendation, referee's report, recommendation form and reference request. Mosel and Goheen (1958a) also include the terms perif and voucher. The structure and format of the reference report can be as diverse as the names by which it is known. It may, for instance, simply be an open-ended request that the writer discuss the applicant. Alternatively, there may be specific guidelines regarding the content of the discussion. Other approaches include rating checklists, forced-choice checklists, and checklists with invitation to comment or embellish on the ratings. One can also consider the telephone check as a form of reference report. Its purpose is essentially the same as other types of reference check but it is distinguished by being verbal in nature rather than written.

Regardless of the terminology that is used, reference reports can be characterised as serving two main functions. The first of these is the verification of self-reported information about an applicant's previous work history. In this case it simply acts as a check on the veracity of the information provided by the applicant. References may also function as assessments of previous work performance or assessments of characteristics thought to be related to work performance, and, as such, can be used to

predict future success on the job. References do not directly sample relevant past behaviours, as advocated by Wernimont and Campbell (1968), but instead rely on judgements and information provided by individuals familiar with the applicant. Jones and Harrison (1982) suggest that the request for reference information may be construed as an attempt to sample indirectly an applicant's past performance. Information gained from the reference report can then be used like any other predictor to help in the selection decision.

For purposes of continuity, the narrative review below will focus on three main areas: (a) surveys examining the use of reference reports, (b) studies investigating the content of reference reports, and (c) research considering psychometric issues associated with the use of references. These areas reflect the major directions that research has taken in this field, and were utilised by Muchinsky (1979) in his critique of the literature. Particular attention will be paid to contemporary studies published subsequent to Muchinsky's review.

Surveys of Use

A recurrent finding from surveys on personnel selection is the popularity enjoyed by reference reports. Surveys have repeatedly shown that reference checks are commonly used during the selection process and that the majority of employers believe they make a valuable and useful contribution. Sleight and Bell (1954) report that nearly two-thirds of the employers they questioned considered letters of recommendation essential, or at least valuable, under certain conditions. Spriegel and James (1958) and Kingston (1971) reported

that more than 66% of the recruiters they surveyed requested reference details or contacted previous employers. The importance attached to referee reports was corroborated in a study by Mosel and Goheen (1958b). They sampled 325 companies to ascertain their practices with regards to reference reports. Over half (51%) of the companies utilised a standardised employment recommendation questionnaire (ERQ) while two-thirds (66.7%) made use of either the ERQ or some other alternative, such as a letter of recommendation or telephone check. It is noteworthy that the majority of the companies that used the ERQ (76%) believed that their selection procedures would be adversely affected should reference information not be available when making hiring decisions.

Spriegel and James (1958) found that nearly 70% of the employers who responded to their questionnaire carried out reference checks on their job applicants. Kingston (1971) also reports very high figures for reference use. In his study, 88% of respondents requested reference details or contacted previous employers. Beason and Belt (1976) note that there was considerable disagreement among the personnel managers whom they surveyed as to the usefulness and value of reference information, but that this was not necessarily reflected in any reduction in the utilisation of reference checks. Almost 30% of the managers they asked said they checked references regularly and 75% indicated that they verified personal references at least occasionally. Additionally, evidence from their survey suggested that many respondents (almost 90%) had a preference for the relatively informal, and perhaps more personal, approach entailed by the telephone check. Sleight

and Bell (1954) reported similar preferences on the part of their respondents. They offer a number of reasons why employers may prefer to use the telephone, including:

1. The belief that previous employers are more likely to be honest if contacted personally, particularly if the recommendation is negative.
2. The idea that personal contact may enable one to obtain additional clues as to the applicant's worth (e.g., via vocal inflection, hesitancies) that would normally be unobtainable through a letter.
3. The fact that it may be quicker than waiting for the return of a letter, hence enabling the employer to make selection decisions more promptly.

More recent surveys confirm that references are still as popular among employers as they have been in the past. Robertson and Makin (1986) found that of the 108 British organisations they sampled, 67% said they always used references during their selection process. Only 4% of the responding organisations indicated that they never used references. Two Australian surveys report very similar results. Patrickson and Haydon (1988) surveyed management selection practices amongst a variety of South Australian business firms and found that reference checks were "usually" or "always" used by at least 70% of the responding organisations. This high frequency of use was consistent for the selection of all levels of manager ranging from first line supervisors through to senior executives. Vaughan and McLean (1989) found that 90% of the business firms they surveyed requested reference reports for at least half the external applications they considered for managerial positions.

The extent to which references are used in New Zealand is more difficult to determine. Information regarding any selection procedure is extremely sparse. However, some impression may be gained from a survey of South Island businesses conducted by Henderson (1987). Over 90% of those responding reported the occasional use of references. Sixty-five percent said that they always used references. The data also showed that references were considered to be quite important in the selection process, and were not being used simply out of convenience. References were ranked third overall as the preferred source of data about applicants; a higher preference was afforded only to the interview and application blank. Mills (1991) in a survey of 30 personnel consulting firms operating in New Zealand found that all of them, without exception, checked references as part of their selection procedures. It is notable that the large majority (97%) of the firms preferred to conduct their reference check by telephone. The importance of information gained from the reference check in the selection procedure as a whole is indicated by the fact that 60% of the firms utilised it to help make decisions regarding the short-listing of applicants. The remaining firms checked references after client organisations had interviewed short-listed applicants, and, in some cases, as a final check after a job offer had been made.

Reference reports also feature quite prominently in the more specialised area of clinical psychology internship selection. Petzel and Berndt (1980) surveyed American Psychological Association approved internship training centres and found that letters of recommendation were ranked as the most important

criterion used by their selection committees. These findings mirror the results from an earlier study by Spitzform and Hamilton (1976) in which they also found letters of recommendation to be deemed the single-most important source of data for intern selection.

Research other than surveys also points to the prominent role reference reports may play during selection. Stedman, Costello, Gaines, Schoenfeld, Loucks, and Burstein (1981) used path analysis procedures to examine the decision making process in the selection of clinical psychology interns. Using an admittedly limited sample (two faculty members) they nonetheless highlighted the critical role of references. For one faculty member, letters of reference directly influenced global ratings and also acted to filter their impression of the applicants' academic preparation. Global ratings in turn, were related to their final ranking of the applicant. For the second faculty member, references had no influence on global ratings of the candidate. However, the analysis did suggest that references were taken into account subsequently, when that faculty member produced a final ranking.

Caution is required when interpreting the results from many of the surveys mentioned. They are often characterised by methodological problems, specifically, very low return rates and non-random sampling. However, despite such difficulties, there is an accumulating mass of evidence implying that references have been, and still are, a very popular method for selecting new employees. It is particularly noteworthy that several studies have shown that references may not only be used frequently, but are also thought to

contribute critical information to the selection process. This makes it all the more alarming that researchers have paid so little attention to references, seemingly content to dismiss them out of hand and ignore them thereafter. Given the central role that references can play in personnel process, one can only reiterate pleas from earlier studies (e.g., Mosel & Goheen, 1958a; Muchinsky, 1979) that more attention should be devoted to this widespread selection technique.

Content of Reference Reports

Little has been written about the information that should be, or is, contained in reference reports. Sleight and Bell (1954) asked 148 employers to indicate the sort of information they would be most desirous of seeing included in the letter of recommendation. The results showed a marked preference on the part of respondents for the reporting of personality characteristics. Personality traits such as responsibility, honesty, cooperativeness, dependability, loyalty, adaptability, and social adjustment were all cited frequently. The employers also indicated that they favoured the inclusion of information concerning the conditions of separation (that is, reasons for their severance from job), and the attitude of the previous employer towards rehiring the applicant.

A later study by Mosel and Goheen (1958b) focusing on companies using ERQs reported somewhat inconsistent results. On the one hand, all of the companies responding to the survey indicated that they requested information

about the applicant's employment history and conditions of separation. In addition, the majority also required some kind of evaluation of the applicant's job performance (e.g., ratings of quality and quantity of work, and attendance). The assessment of personality characteristics was apparently uncommon. On the other hand, and somewhat surprisingly given the lack of emphasis in reporting such information, personality characteristics of one type or another (e.g., honesty, cooperativeness, character) were judged the second most useful ERQ items when making a hiring decision.

Peres and Garcia (1962) reviewed the content of 625 reference letters for engineering applicants at a large nuclear research and development facility. They found that the applicants were usually described using adjectives or descriptors reflecting personality dimensions and traits rather than observable behaviours. Using content analysis, they were able to extract 170 different adjectives from the letters. The authors then asked 100 technical supervisors to rate (on a five-point scale) how applicable each adjective was in describing their "best" engineer. An equivalent number of supervisors were required to rate the applicability of each adjective in describing their "poorest" engineer. The adjectives were then factor analysed and the factor scores correlated with the supervisors' ratings. Five critical dimensions were identified: cooperation-consideration, mental agility, urbanity, vigour, and dependability-reliability. They also found that adjectives illustrative of mental agility tended to be the most discriminative of good versus poor engineers, as determined by supervisors' ratings. Adjectives illustrative of cooperation-consideration were least discriminative. Peres and Garcia suggest that

referees asked to write a recommendation for an applicant they do not truly believe to be qualified for the position may often resort to describing the individual as a "nice person". In other words, such referees may use adjectives common to the least discriminative factors (e.g., good-natured, likable, friendly, talkative, bold, sociable, etc.), what Peres and Garcia termed "damming with faint praise".

The desirable content of letters of recommendation can also be partly inferred from studies investigating other phenomena, such as reliability or validity. For example, Baxter, Brock, Hill, and Rozelle (1981) report an average of nearly eight psychological qualities (e.g., intelligent, reserved, not unimaginative) attributed to, or disassociated from, their student applicants in each letter of recommendation. The average number of descriptive characteristics (e.g., sex, age, ethnicity) recorded for each applicant was only 1.6. Browning (1968) used a reference form in his study on the validity of reference ratings for teachers that assessed the following factors: teaching ability, relationship with children, professional relationships, community relationships, personal qualities and health. The reference reports used in Jones and Harrison's (1982) study of Royal Naval Officer trainees included a wide range of information covering factors such as each candidate's academic performance, extra-curricular activities, positions of responsibility and aspects of character and behaviour. Clearly, the information demanded by employers in reference reports can be quite eclectic.

Knouse (1983) examined the relative impact of specificity and favourability of information contained in letters of recommendation on personnel managers' perceptions of the reference report, the referee, and the applicant. Using a 2 x 2 x 2 factorial design, he varied the specificity of information (specific examples versus no examples), the presence of numerical data (numerical data versus nonspecific adjective modifiers), and favourability (favourable letter versus one unfavourable statement). Readers perceived the example-specific variation as providing more information, being more favourable, showing the applicant to be a better potential manager, demonstrating that the writer was more familiar with the applicant and being better at writing than the no specific example variation. More to the point, the managers indicated that they would be more prepared to hire the candidate whose letter of recommendation included specific examples. Strong effects were also found for favourability. Not surprisingly, managers perceived the favourable letter variation as being more favourable to the applicant. However, they also believed the applicant to be a better leader, better potential manager and better overall than was the case with the variation containing one unfavourable statement. No main effects for numerical specificity were found, although this factor did interact with example specificity. When examples were present, numbers slightly decreased perceptions, and when examples were absent, numbers enhanced perceptions. An interaction effect between favourability and example specificity was also discovered. When examples were included, the impact of an unfavourable statement was lessened, with managers perceiving such letters as containing more information and the writer as being more credible.

From the few research studies that have been conducted, it is plain that the content of reference reports is far from standardised. Most studies have indicated a strong preference for information about personal characteristics and reasons for leaving the position. More recently, the study by Knouse (1983) has shown that the favourability of reference reports influences managers' perceptions of ratees' performance, and that with the inclusion of specific examples they were more prepared to extend an offer of employment to the applicant.

Psychometric Issues

Reviewers (e.g., Muchinsky, 1979; Reilly & Chao, 1982) have attributed the scant attention reference reports have received in the literature to their well-documented deficiencies in regard to reliability and validity, and their problems with leniency and restriction of range. Early studies by Mosel and Goheen (1952; 1958a; 1958b; 1959) and Goheen and Mosel (1959) are widely cited as indicative of such deficiencies. More recent research continues to highlight problems associated with the use of reference reports, as evidenced by the following review.

Reliability

Studies investigating the reliability of references are few and far between. One of the earliest is reported by Mosel and Goheen (1952). They examined the

reliability of ratings of job applicants from different reference sources. Using an ERQ, they gathered 2,500 references on 904 applicants for nine different civil service jobs. Each recommendation (ERQ) required the respondent to rate the applicant on a variety of skills, knowledge and personality traits deemed to be related to job success. Depending on the particular job, applicants could be rated on as few as 5 to as many as 39 different factors. The agreement between respondents rating the same applicant on each of these different factors showed considerable variability (r = .01 to .98) but was generally very low. In fact, 80% of the reliability coefficients were below .40. Mosel and Goheen concluded that the reliability of the different rating factors was related to some extent to the type of job. However, they also point out that a wide range of reliabilities existed within the questionnaire for any single job, and that even the better questionnaires contained enough unreliable items to warrant improvement. One may add that such unreliability would make any interpretation of the ERQs and consequential selection decision extremely difficult.

A subsequent study by Mosel and Goheen (1959) also considered the reliability of different reference sources. They analyzed over 3,000 ERQs provided for applicants for seven different jobs in government service. Reference sources were categorised into five major groups: previous employers, co-workers, subordinates, teachers and acquaintances. They then compared the mean ratings provided from each reference source for each of the positions. The results showed a tendency for acquaintances to be the most favourable in their ratings, followed by subordinates and co-workers.

The ratings from teachers were inconsistent and those from previous employers were generally the most severe. A degree of circumspection is required when interpreting their results. Differences in mean ratings were not large and were not compared statistically, so may reflect chance fluctuations. However, the evidence suggests that alternate reference sources may provide different evaluations of applicants.

While the ratings of applicants may vary depending upon the nature of the respondent, it is also feasible that the overall ranking of the applicants may remain constant across each type of respondent. Mosel and Goheen (1959) investigated this possibility by considering the ERQs submitted on 116 government printers from various reference sources. The results showed little consensus between the different respondents. ERQs submitted by previous supervisors correlated r = -.12 with those provided by acquaintances. The averaged combined ratings from supervisors and co-workers correlated .24 with judgements by acquaintances. Although the latter result was significant, the level of agreement is still very low. Once again, caution must be advocated in interpreting these figures. Mosel and Goheen appeared to analyze only a portion of the available data, ignoring the six other jobs for which ERQs were collected. The basis for restricting the analysis to printers (the only non-professional job in the sample) is not made clear in their report. Furthermore, the analysis of the reliability of ERQs provided for printers also appeared incomplete. Figures documenting the agreement between supervisors and co-workers and between co-workers and acquaintances were not provided. The reason for their omission from the report is not readily

apparent. Finally, as Muchinsky (1979) suggests, it may not be reasonable to expect high levels of agreement across different types of raters as they may be basing their evaluations on divergent sources of knowledge regarding the applicant. The key issue then becomes one of deciding which reference source is most valid for selecting applicants for particular jobs in the work force.

It certainly may not be appropriate to expect high levels of inter-rater reliability in the case of different reference sources. However, it is reasonable to expect at least moderate levels of agreement in the case of referees who occupy similar positions and are evaluating the same applicant. Unfortunately, the results from the study by Baxter et al. (1981) suggest that reliability may be poor even under those circumstances. They examined letters of recommendation on candidates for admission to graduate school. Ratings by a panel of judges showed that there was very little agreement between letters of recommendation from different referees regarding the same student (low reliability). Indeed, there was more consistency or agreement between letters from the same referee concerning different students. Baxter et al. concluded that the letters of recommendation in their study were nondiscriminative, nonconsensual and nondifferentiating. Furthermore, post hoc analysis of a subsample of the letters revealed that the results were not attributable to any differences in context or time of acquaintance on the part of the writers. In fact, the letters became progressively less discriminative and consensual as the period of acquaintance between the writer and the applicant increased. It is possible that the narrative format of the letters may

have contributed to their low reliability. The lack of focus such a format entails may have exaggerated the impact of different literary styles, modes of expression and so forth. This in turn may have encouraged writers to set out their unique and individualised impressions. More mundanely, Baxter et al. report only moderate levels of agreement amongst their expert judges who were rating the reference letters. Reliability ranged from r = .55 to .79. Clearly, low levels of agreement could impact significantly on the results and at the very least would make it difficult to draw any firm conclusions.

Some evidence of intrarater consistency is provided in a study by Ceci and Peters (1984). The focus of their naturalistic study was on the effects of confidentiality on letters of recommendation. However, the experimental design also enabled an evaluation of the reliability of ratings provided by a referee for the same applicant on more than one occasion (a type of test-retest reliability). The authors report no significant differences in ratings in two letters concerning the same applicant sent to different universities except when one of the letters was thought to be confidential. Although the evidence to date on the reliability of reference reports is a little troubling, it is by no means conclusive. More studies need to be conducted and it would be fruitful for investigators to explore the conditions and possibilities for improving reliability.

Validity

Although research on the validity of references has generally been quite sporadic and haphazard, it can be characterised as falling into five major categories. These include: studies that have examined the validity of references as a function of the source of the reference; studies that have evaluated the validity of references for different target populations; studies that have assessed the validity of references in relation to different formats and content; studies that have examined miscellaneous threats to the validity of references; and finally, meta-analytic studies that have reviewed previous research on the validity of references.

Source of the Recommendation

Kornhauser (1927) looked at the validity of recommendations for predicting the academic success of college students. He found that references from previous employers and teachers correlated most highly with later academic success (r = .35 and .26 respectively). Recommendations from friends, lawyers, ministers and others were not significantly related to the criterion outcome.

Mosel and Goheen (1959) conducted a similar study in which they compared the validity of ERQs submitted by different referees for applicants for five different trade jobs in the federal service. The data consisted of 795 ERQs on 400 civil service incumbents. ERQs were submitted by one of five different categories of referee. These were personnel officers, supervisors, co-workers,

acquaintances and relatives. The criterion of job success was the summated score from a 10-factor performance rating from the present supervisor. Somewhat surprisingly, ratings from acquaintances bore the strongest relationship to the criterion (r = .20). The validity coefficients for the remaining referees ranged from .19 for previous supervisors to -.16 for relatives.

Browning (1968) evaluated the validity of different reference sources for predicting the performance of newly employed teachers. A standardised reference form consisting of six 4-point adjectival scales was used for all respondents. At the end of their first year of employment the teachers were rated for competency by their local principal. These ratings were used as the criterion measure. In all, 2,221 ratings from 11 different referee types (an average of 4.4 references per teacher) were used to compute the validity coefficients. Ratings by the previous supervisor were most predictive of performance (r = .23). Thereafter they ranged in value from .22 (other superintendent) to -.03 (professor of practice teaching). The average validity across the 11 different referee types was only .13.

Overall, the validity of references for predicting subsequent performance on the job was found to be poor. The three studies also show that different classes of referee are liable to achieve varying levels of predictability. Previous employers/supervisors appear to be the best source of reference ratings. Muchinsky (1979) suggests their superiority may be attributed to their greater knowledge of the applicant's work performance, an idea congruent with the

concept of behavioural consistency. Although supervisors were generally superior to other reference sources, it must be remembered that the level of validity they achieved was still only modest at best.

Target Population

Only one study has set out to determine the validity of references for different groups of employees. Mosel and Goheen (1958a) collected ERQs for 1,193 civil service employees in 12 skilled occupations. Average scores for each ERQ were correlated with performance ratings provided by each participant's current supervisor. The validity coefficients showed considerable variability, ranging in value from r = .29 to -.10, with an average of .12. Subsequent examination of the data revealed that several of the ERQ items had very low discrimination power. This prompted Mosel and Goheen to re-analyse their data using only ratings of occupational ability and ratings of character and reputation. Variability in the validity coefficients declined marginally compared to the overall analysis. The average validity coefficient increased when ratings of character were used but decreased slightly when ratings of occupational ability were utilised. Mosel and Goheen concluded that ERQs for the trades' occupations appeared to have little value in predicting subsequent job performance. Validity coefficients varied from job to job, but it is not possible to say from their study if this variation reflects systematic differences in the predictability of references for different occupations.

Format and Content

One of the most frequently cited problems of references is their proneness to leniency errors and consequential restriction in range that results. One approach taken by researchers to minimise this tendency towards excessive leniency on the part of referees is to adopt an alternative format for reference reports. The most popular of these has been the forced-choice rating scale which is specifically designed to reduce leniency in ratings.

Newman and Howell (1961) examined the validity of forced-choice items for the prediction of the performance of medical officers. They correlated scores from a forced-choice rating form with three criterion measures: ratings from an officer efficiency report, and evaluations collected from work associates assessing performance and personality. The authors report validity coefficients of r = .27, .29 and .35 for the three criterion measures respectively.

Rhea, Rimland, and Githens (1965) compared a forced-choice reference form and a checklist reference form for predicting performance ratings of junior naval officers. They reported that 17 out of 36 validity coefficients were significant for the forced-choice form while all 42 of the possible validity coefficients were significant for the standard checklist reference form. A follow-up study (Rhea, 1966) compared the same two types of reference form for the prediction of performance of potential naval officers attending officer candidate school. In this case the average validity of the forced-choice

reference form was .20 while the average validity for the checklist was .10, effectively reversing the results from the first study.

Carroll and Nash (1972) developed a forced-choice reference check for clerical workers. They derived 24 dyads consisting of behavioural statements used to describe clerical workers. Items were equated with respect to frequency of use and favourability, but differed in their ability to predict job success. The reference form was then sent to the former employers of clerical workers. Scores on the reference check correlated r = .21 with a composite performance rating completed by current supervisors after four months on the job. The reference check was also found to be somewhat predictive of turnover, correlating -.18 with terminations. In a post-hoc analysis Carroll and Nash divided their sample into a validation group and a cross-validation group. They then identified five moderating variables affecting the validity of the reference check. From this they developed several decision rules to enhance predictability. Application of the decision rules to the cross-validation group reduced the group size by 31%, but resulted in a validity coefficient of .47. Subsequent item analysis of the reference form and use of a scoring key containing only the 12 best items further improved the validity of the reference check for the cross-validation group to .56.

Bartlett and Goldstein (1976; cited in Reilly & Chao, 1982) eschewed the written reference form and instead examined the validity of a telephone reference check for predicting turnover. Using support personnel from a large technical organisation, they obtained a validity coefficient of .27 for the

prediction of involuntary termination and .07 for all terminations. (Maximum possible validity coefficients obtainable were .38 and .25 respectively due to the dichotomous nature of the predictor data and extreme splits that occurred.)

Evidence regarding the influence of the content of reference reports on validity is sparse. No studies have directly assessed the impact of this variable in terms of predicting job success. However, some indication of the possible effects of content may be gleaned from the previously described study conducted by Knouse (1983; p. 9). The Knouse study clearly demonstrates that even relatively minor variations in the content of letters of recommendation can exert a significant influence on readers' perceptions and, hence, quite possibly, their subsequent decision to hire. An alternative approach to the issue of content revolves around the question of what information results in the most valid appraisals of applicants. It is difficult to draw any definite conclusions because this problem has not been explicitly addressed. Some indirect evidence is available from studies that have investigated other aspects of validity. Mosel and Goheen (1958a) presented the validity coefficients for separate items drawn from the ERQ used for employees in the skilled trades. Ratings of character and reputation resulted in a higher mean validity coefficient than ratings of occupational ability or overall mean ERQ scores. Browning (1968) found that a composite rating was superior to ratings on any individual factor for the prediction of the performance of newly employed teachers. The results from both of these

studies are difficult to interpret since they did not control for the source of the reference.

Jones and Harrison (1982) examined the usefulness of reference reports for the prediction of Royal Naval Officer Cadet training success. The reference reports consisted of ratings by headteachers on seven 5-point scales assessing a variety of factors including: application to studies, involvement in sporting and extracurricular activities, discharge of responsibility, character, relations with fellows, influence on and leadership of fellows and overall contribution to the school. A composite rating, the sum of all seven scales, was also calculated. Criterion data consisted of an examination (professional mark) and an assessment of leadership and conduct (former service mark). Scores on both criteria were also summed to produce an overall grade (total mark) from which the class pass could be established. After correcting for restriction in range, the authors found that the summated rating was the most predictive of cadet success, correlating .36 with the outcome measure. No clear superiority for any individual scale was established. Jones and Harrison suggest that their generally improved validity coefficients reflect the origin of the reference information used in their study. They argue that teachers are familiar with evaluating students, and unlike other referees may experience less pressure or expectation to provide a biased appraisal. Furthermore, they also suggest that the teachers acting as referees would have had greater opportunity to observe the subjects of their reports than may be the case with other referees. However, these explanations remain contentious for two reasons. Firstly, one can query how often headteachers

are provided with the opportunity to observe their pupils closely, particularly if their own job involves a heavy administrative workload. Secondly, an examination of the means and standard deviations of the ratings provided by the headteachers reveals considerable skewness. Whether this was due to leniency error or restriction in range is, as Jones and Harrison note, open to interpretation. It seems premature at this stage to attribute any improved results to the source of the report. Rather, further investigation of this variable appears warranted.

Miscellaneous Threats to Validity

There have been several studies focusing on other variables that could influence the validity of references. Kryger and Shikiar (1978) considered the possibility of sexual discrimination in the use of letters of recommendation. They sent one of eight different versions of a letter of recommendation (male or female writer, male or female applicant, favourable or unfavourable letter) to be evaluated by 128 male personnel managers. They hypothesised that female applicants would be less favourably evaluated than male applicants and that applicants with letters of recommendation written by females would be less favourably evaluated than applicants with letters written by males. Contrary to their expectations they found that female applicants were actually preferred to male applicants in terms of proceeding with an interview. The personnel managers also judged the female applicants to have more initiative and responsibility and a greater capacity to learn quickly than the male applicants. The sex of the writer had no significant main effects on the

dependent variables but did interact with letter favourability, so that a woman writing a favourable letter was judged to like the applicant significantly more than a woman writing an unfavourable letter. There was no effect for males regardless of whether they wrote a favourable or unfavourable letter. Sex of the letter writer also entered into a higher order interaction. Female writers were seen as requiring significantly greater feedback when they favourably evaluated a female applicant who subsequently performed poorly compared to a male applicant. Similarly, a female writer unfavourably evaluating a male applicant who subsequently performed well was seen as warranting significantly greater feedback than if the same error had occurred in evaluating a female applicant. For male writers there were no significant differences.

Shaffer and associates (Shaffer, Mays, & Etheridge, 1976; Shaffer & Tomarelli, 1981) have investigated the impact of confidential versus nonconfidential letters of recommendation on hiring decisions. In a simulated employment study, Shaffer et al. manipulated applicant competence, enthusiasm of the letter, sex of reader and choice of placement file selected by the applicant. They found that their hypothetical employers expressed a preference for the job candidacy of applicants who selected confidential rather than open placement files. The effect was found at each level of applicant competence, letter enthusiasm and regardless of the sex of the reader. The preference for applicants with confidential files was not moderated by information that established the applicant's competence independently of the letter of recommendation. Furthermore, confidential file placement seemed to exert a

kind of halo effect in that participants indicated that such applicants were somewhat more socially attractive and preferable as job supervisors. In a follow-up, Shaffer and Tomarelli carried out an archival study to assess the influence of confidentiality in a "real-life" rather than hypothetical selection situation. Using data from 253 applicants for graduate studies in psychology at the same university, they found that graduate admissions officers favoured applicants whose letters of recommendation were confidential. The effect of confidentiality was once again found at each level of academic competence and in addition across each of the four doctoral programmes for which applications were received.

Ceci and Peters (1984) examined the extent to which the perceived confidentiality or nonconfidentiality of a letter of recommendation influenced the letter writer's evaluation of prospective graduate school applicants. Faculty advisers were approached by student confederates over a period of three consecutive months and asked to complete standard letter of recommendation forms supposedly to be sent to three different universities. The forms were identical apart from the university mastheads. Two of the three letter of recommendation forms were marked the same (either confidential or nonconfidential) while the remaining form was oppositely marked. Half of the students initially requested a confidential letter and half requested a nonconfidential letter. The order in which the three different university letters were requested was also counterbalanced with the confidential and nonconfidential instructions. The dependent variable consisted of ratings of the students on the 10 scales contained in the

recommendation forms. Analysis of the results revealed significant differences between the confidential and nonconfidential letters with students being rated lower on most scales in the confidential letter. Taken together, the results of these studies strongly suggest that the confidentiality or nonconfidentiality of a letter of recommendation can influence both the reader and the writer of such letters.

Paunonen, Jackson, and Oberman (1987) used simulated employment interviews to investigate the effects of applicant personality attributes and the letter of reference on personnel selection decisions. The hypothetical applicant's personality attributes were varied so as to be characteristic or not characteristic of incumbents in the job. Each participant studied a job description, read a letter of recommendation and listened to taped segments from a job interview during which the applicant divulged personality information only. They found the perceived competence of the job applicant established in the letter of reference had a powerful effect on their student judges' ratings of suitability, more competent applicants being rated as more suitable, which masked the effects of the personality manipulation established in the simulated interview. In a subsequent experiment, where the levels of competence portrayed in the reference letter were less discrepant and the personality manipulation in the interview was made more salient by presenting opposing rather than unrelated personality types, the main effect for the competence manipulation was still substantial. The study is admittedly of limited generalizability because of its use of simulated interviews and student participants. However, the results do suggest that the

information contained in reference letters may exert a substantial impact on an individual's decision to hire.

Meta-Analytic Studies

Reilly and Chao (1982) reviewed the validity of a variety of alternative employee selection procedures. Using a total of seven studies they estimated a validity coefficient of .18 for reference checks predicting rating criteria and .08 for the prediction of turnover. They concluded that the utility of reference checks was limited and recommended that they generally not be used in the selection process.

Hunter and Hunter (1984) also reviewed alternative predictors of job performance. Unlike Reilly and Chao (1982), who simply averaged correlation coefficients across studies, they utilised more sophisticated formulae able to correct for sampling error, error of measurement and range restriction (Hunter, Schmidt, & Jackson, 1982). The estimated corrected mean validity coefficient was .26 using supervisors' ratings as the criterion. For the prediction of tenure, the average validity coefficient was .27 which surpassed all other alternative predictors considered by Hunter and Hunter, including biographical data. However, it should be borne in mind that the prediction of tenure for reference checks was based on only two correlation coefficients.

The Meta-Analysis

It has already been noted that there is a paucity of research on reference reports. Particularly disturbing is the marked absence of any contemporary studies; most of the research that has been published is now quite dated. This is in stark contrast to research on the interview. Like reference reports, the interview is a very popular selection technique that traditionally has not fared well in empirical assessments of its reliability and validity. Until recently, reviewers have been virtually unanimous in their condemnation of the interview (e.g., Arvey, 1979; Arvey & Campion, 1982; Mayfield, 1964; Reilly & Chao, 1982; Ulrich & Trumbo, 1965; Wagner, 1949; Wright, 1969). However, unlike reference reports, this has not deterred researchers from continuing to investigate, experiment, and publish material on the interview. Such assiduousness in the face of contradictory evidence appears to have paid dividends in the last few years. Several researchers have shown that the interview can achieve quite acceptable levels of validity (e.g., Arvey, Miller, Gould, & Burch, 1987; Campion, Pursell, & Brown, 1988; Latham, Saari, Pursell, & Campion, 1980; Latham & Saari, 1984; Marchese & Muchinsky, 1993; Weekley & Gier, 1987; Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989). Inevitably, one has to ask why reference reports have been ignored and whether they merit such perfunctory treatment. Perhaps all that is required is further research investigating their potential and, like the interview, they will also evidence improved validity? Muchinsky (1979) made a similar plea at the end of his review when he asked researchers to take a

more innovative approach in the development and evaluation of reference reports.

The focus of the first part of the present study is a meta-analytic review of reference reports. While the Hunter and Hunter (1984) review was broad in scope, examining as it did many different selection procedures, their data base for reference reports consisted mainly of the studies reviewed by Reilly and Chao (1982). Given the persistent popularity of reference reports, the advent of recent studies highlighting their role in the selection process and their potential for improved validity, a comprehensive meta-analytic review focusing on reference reports and drawing on as large a data base as possible seems timely. Such an approach will enable a more accurate estimate of the true validity of reference reports, and also allow for moderator analyses should they prove necessary. The following chapter describes the literature search, decision rules, and meta-analytic method used for the suggested quantitative review of reference reports.

Chapter 3

Meta-Analysis

The present chapter focusses on the meta-analytic review of the predictive validity of reference reports. The chapter begins by describing the method used and then the results from the analyses are reported and discussed. Finally, links between reference reports and performance rating research are introduced.

Method

Literature Search

The first step taken was a thorough review of the published literature dealing with the validity of reference reports. Appropriate bibliographic indexes were consulted to track down relevant studies (e.g., Psychological Abstracts, Social Sciences Citation Index) and a computerised search of on-line data bases was carried out (Psych-Info, NTIS, Social Sciences Citation Index, Dissertation Abstracts, ABI/Inform) using the DIALOG Information Retrieval Service. Particular attention was paid to the reference sections from the collected reports and from previously published review articles so that unpublished studies, and any early publications that were not included on the computerised data base, could be located.

The search procedure yielded eight studies, six published and two unpublished, from which 125 usable validity coefficients were derived. However, 62 of the 125 validity coefficients were obtained from the two unpublished studies. Table 1 shows the number of validity coefficients and total sample size contributed by each study in the meta-analysis. No study was excluded from the analysis on any a priori basis.

Table 1
Studies, sample type, total number of participants and validity coefficients contributed by each study to the meta-analysis

Study	Sample Type	Total Number of Participants	Total Number of Validity Coefficients
Mosel & Goheen (1958a)	Skilled	3122	34
Mosel & Goheen (1959)	Skilled trades	1143	8
Newman & Howell (1961)	Medical officers	384	1
Rhea, Rimland, & Githens (1965)	Naval officer cadets	26162	31
Rhea (1966)	Naval officer cadets	18211	31
Browning (1968)	Teachers	2221	11
Carroll & Nash (1972)	Clerical	94	1
Jones & Harrison (1982)	Naval officer cadets	824	8

Decision Rules for Coding Studies into the Meta-Analysis

Certain decision rules pertaining to the nature of the validity data that should be recorded were established based largely on recommendations from Hunter

and Schmidt (1990). They note that it is frequently possible to derive more than one estimate of predictor-criterion relationships from within the same study. This presents no difficulties for the meta-analyst if the data are statistically independent. In such cases the cumulation process can proceed as if the values were from different studies. Hunter and Schmidt refer to such cases as fully replicated designs.

More problematic is the case of conceptual replication. Such a replication occurs when more than one observation pertinent to a specific relationship is derived for each participant. According to Hunter and Schmidt (1990), the most common instances of this are the use of multiple measures to assess a given variable, and assessment across multiple settings. Any assumption of statistical independence is likely to be compromised in most conceptual replications. Statistical independence can only be assured if the correlations between the multiple measures, or across multiple settings, are all zero. If a meta-analysis includes groups of correlations from the same study samples and cannot guarantee the statistical independence of the data, then it is probable that the sampling error variance will be underestimated. Consequently, there will be an undercorrection for sampling error which means that the final estimate of the population variance, S_r^2, will be too large. The result is likely to be an underestimate of the degree of agreement across studies encouraging what is likely to be a fruitless search for moderator variables. Hunter and Schmidt go on to point out that if the number of correlations contributed by each study is few relative to the total number of correlations in the meta-analysis, then any resulting cumulation will contain

very little error. On the other hand, if a large number of values are contributed from any one study then difficulties in interpreting the results may occur.

It is noteworthy that for the present meta-analysis one sample of participants would have contributed 78% of all the validity coefficients if each conceptual replication was included in the analysis. Because of this problem, it was necessary to develop a set of decision rules for including validity coefficients in the analysis.

First, in situations where validity was assessed for two or more predictors of the same type for a given sample, each coefficient was recorded separately. This rule was relevant to the studies by Rhea et al. (1965) and Rhea (1966) where a forced-choice recommendation form was compared to other more traditional formats. The rule was also applied in cases where different referees supplied judgements for the same group of participants or sub-sample of participants (e.g., Browning, 1968; Mosel & Goheen, 1959). The application of this particular decision rule helped minimize the potential loss of information from a strict adherence to the principle of statistical independence.

Second, for those studies that included validity coefficients for several dimensions of a particular criterion measure (e.g., performance ratings, class ranking, ratings of quality and quantity of work), only the coefficient for the overall or summary measure was recorded, if it was available. For cases

where this overall estimate was not available, the validity coefficients for the various dimensions were averaged and that figure recorded for the analysis. This approach was adopted by Hunter and Schmidt (1990) for their own studies and helps ensure statistical independence between criterion measures.

Third, in those cases where both uncorrected and corrected validity coefficients were reported, only the unattenuated figures were included in the analysis. Corrections were performed as part of the meta-analysis.

Fourth, predictor-criterion relationships for different groups of referees combined were only included if no sub-group analysis was reported in the study.

In excess of 500 validity coefficients were reported in the eight studies comprising the meta-analysis sample. Application of the decision rules reduced the number of usable validity coefficients to 125. The utilization of the second decision rule was responsible for the vast majority of eliminations from the data pool. Figure 1 depicts the distribution of validity coefficients used for the meta-analysis.

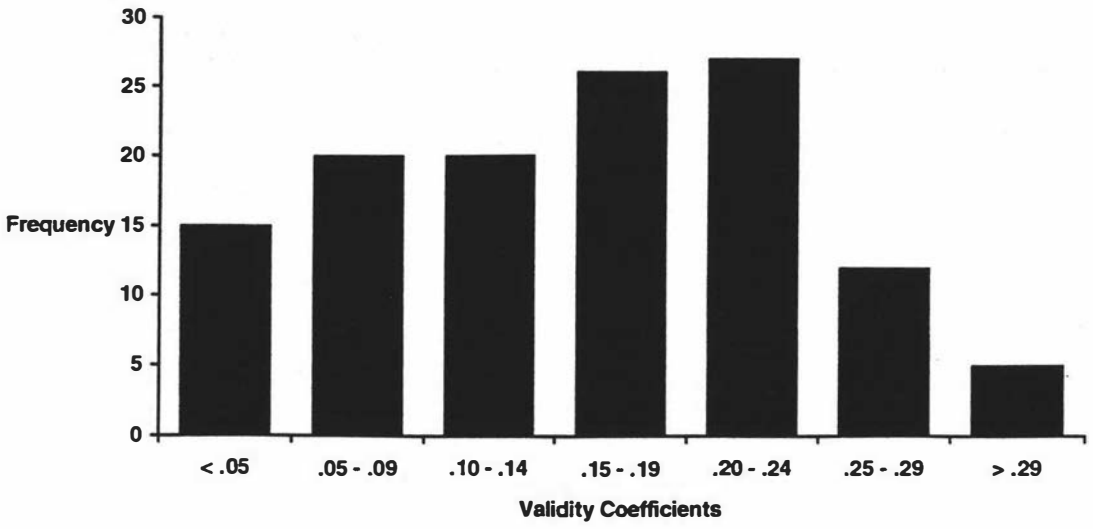


Figure 1
Distribution of validity coefficients from studies in the meta-analysis

Meta-Analysis Method Used

The new meta-analytic formulae proposed by Raju et al. (1991) were used to analyze the sample data. Monte Carlo studies completed by Raju et al. suggest that the new approach is generally more accurate in estimating the mean and variance of true validities than other existing correlation-based procedures. The new technique can account for the sampling error in artifacts (predictor and criterion reliability only) and can readily incorporate varying proportions of sample-based artifact data should they be available. Even in instances where hypothetical artifact distributions are utilised, as is the case for the present analysis, the new approach is easier to apply as it requires only the mean values for the appropriate artifact distributions rather

than the means and variances that are necessary for traditional meta-analytic procedures.

The Raju et al. (1991) method entails calculating an unattenuated value for each validity coefficient ($\hat{\rho}_i$) using the following formula;

$$\hat{\rho}_i = \frac{k_i\, r_i}{\sqrt{r_{x_i x_i}\, r_{y_i y_i} - r_i^2 + k_i^2\, r_i^2}} \qquad [1]$$

$$k_i = 1/u_i$$

where r_i represents the correlation between the predictor (x) and the criterion (y) in a sample from population i or for validity study i; $r_{x_i x_i}$ = the sample based, attenuated reliability for the predictor; $r_{y_i y_i}$ = the sample based, attenuated reliability for the criterion; u_i represents the ratio of restricted standard deviation to unrestricted standard deviation on the predictor.

The average true validity, M_p, can then be estimated as the sample size weighted average of $\hat{\rho}_i$. In order to establish the true variance, V_p, it is necessary to compute the sampling variance, V_{e_i}, of each unattenuated coefficient, $\hat{\rho}_i$. For the current analysis equation 2 was used. It should be noted that this formula assumes that range restriction, criterion reliability

and predictor reliability are all fixed. This is the case, as with the present analysis, when artifact values are used.

$$\hat{V}_{e_i} = \frac{k_i^2 \, r_{x_i x_i}^2 \, r_{y_i y_i}^2 \left(1 - r_{x_i y_i}^2\right)^2}{N_i \, \hat{W}_i^3} \qquad [2]$$

where $\hat{W}_i = r_{x_i x_i} \, r_{y_i y_i} - r_i^2 + k_i^2 \, r_i^2$

A sample size weighted sampling variance is then calculated, using equation 3, to provide an estimate of the overall sampling variance, V_e.

$$V_e = \frac{N_1 \, V_{e_1} + \ldots\ldots\ldots + N_m \, V_{e_m}}{\sum_{i=1}^{m} N_i} \qquad (3)$$

Using equation 4, the population variance, V_p, can be estimated by subtracting the overall sampling variance, V_e, from the sample size weighted variance for the unattenuated coefficients, $V_{\hat{p}}$.

$$V_p = V_{\hat{p}} - V_e \qquad (4)$$

For all analyses, mean artifact values posited by Pearlman, Schmidt, and Hunter (1980) provided the basis for corrections to the validity coefficients for

range restriction and criterion unreliability. The studies contributing to the meta-analysis provided only sporadic information on likely artifact values, necessitating the use of hypothetical artifact values (two reports, from the same study, of criterion reliability averaging r = .89; two estimates of range restriction averaging u = .48). Consistent with other meta-analytic studies (e.g. Marchese & Muchinsky, 1993; Wiesner & Cronshaw, 1988) no corrections for attenuation due to predictor unreliability were carried out. Although an argument can be made for correcting for predictor unreliability (see Hunter & Schmidt, 1990) such corrections are not normally recommended for meta-analytic studies in the field of personnel selection (see Schmidt, Hunter, Pearlman, & Hirsh, 1985) where practitioners must rely on the fallible instruments available to them. Validity coefficients were corrected using training and proficiency criterion reliability values where it was appropriate.

Chi-square values for significant variation across studies were computed to detect potential moderator variables. If the chi-square test was significant, the credibility interval included zero, and if the residual variance remaining after corrections for sampling error exceeded 25%, then the effects of likely moderators were assessed by conducting meta-analyses on sub-sets of the validity coefficients sorted according to the potential moderator variable thought to be operating (see Whitener, 1990).

For all analyses the following figures were calculated and reported. The total sample size, the mean validity coefficient corrected for sampling error only

(bare bones analysis), the fully corrected mean validity coefficient, the sample weighted variance for the corrected correlations, the sample weighted sampling error variance, the percent of variance accounted for by sampling error, the 95% credibility interval and the chi-square value for the test of variation in correlations.

Results

The results of the meta-analysis incorporating all the validity coefficients is presented in Table 2. The mean validity coefficient of .15 from the bare bones analysis more than doubled following corrections for range restriction and criterion unreliability (r = .31). Sampling error was found to account for only 44% of the variance in the sample correlations. Hunter and Schmidt (1990) and others (e.g., Tukey, 1960) have alerted researchers to the potentially biasing effect of outliers on estimations of means and standard deviations. Wright et al. (1989) in their meta-analysis of structured interviews found one outlier study exerted an unwarranted impact on the results of their analysis. Replication of the meta-analysis by Wright et al. with the outlier study removed resulted in a substantially reduced variance among the correlations, eliminating the requirement to identify a moderator variable.

A re-examination of the studies comprising the sample in the present analysis revealed several extreme coefficients. Considering that there was potential for

aberrant values to influence the results, it was deemed prudent to replicate the meta-analysis with any such outlier values excluded. To reduce the influence of outliers, 5% of the most extreme values in the sample were eliminated, with the caveat that they must, in addition, fall more than three standard deviations from the mean.

Table 2
Meta-analysis of the predictive validity of reference reports

	All Studies	Outliers Excluded
Total sample size[a]	52161 (125)	51687 (119)
Mean validity coefficient[b]	.15 (.31)	.15 (.32)
Sample weighted variance in corrected correlations	.019	.017
Sample weighted sampling error variance	.008	.008
% variance accounted for by sampling error	44	47
95% credibility interval	.10 to .52	.13 to .51
χ^2	296.9*	252.9*

[a] The number of coefficients contributing to the total sample is given in parentheses.
[b] Means were corrected for restriction of range and criterion unreliability; corrected values are given in parentheses.
* $p < .0005$.

The results of the analysis without outliers is shown in the second column of Table 2. The significant chi-square value ($\chi^2 = 252.9$, $df = 118$, $p < .0005$) continued to imply the presence of a moderator variable. Hunter and Schmidt (1990) have cautioned meta-analytic researchers about relying unduly on the chi-square test for the detection of non-trivial variation in their results. However, for the current analysis the remaining variance unaccounted for by statistical artifacts exceeded 50%. Application of Hunter and Schmidt's

(1977) "75% rule" and the results from the chi-square test suggests one or more moderator variables may be present.

Post hoc Analyses

Several post hoc analyses of the data were conducted to ascertain and delineate the impact of potential moderator variables. Wiesner and Cronshaw (1988) found that interview structure moderated the predictive validity of interviews. A categorisation based on similar conceptual principles seemed pertinent to the present study given that several authors (Carroll & Nash, 1972; Muchinsky, 1979) had already intimated that a more structured approach, namely forced-choice, could surpass other forms of reference report in terms of predictive validity. In the present case, it was decided to group the data according to the efforts made on the part of the authors to use established psychometric procedures in the development and construction of the reference forms. Particular attention was paid to any attempts to combat leniency errors as these have been identified as a major problem in this area (e.g., Muchinsky, 1979). The following categories were derived:

1. Highly Structured - reference forms attempting to incorporate sound psychometric principles. These included such elements as the use of job analysis information, the use of multiple items and summated scores, item analysis and, in particular, specific attempts to combat problems of leniency, halo, and bias. (E.g., studies using forced-choice reference forms.)

2. Structured - this category included those studies that attempted to collect ratings in a systematic manner using measures assessing factors thought to be related to job performance. Scores on these measures are derived by objectively combining item or scale scores. No special methods are used to combat problems of leniency, halo or bias. (E.g., studies using Likert scales measuring multiple dimensions and reporting summated scores.)

3. Semi-structured - this category comprised attempts to measure an individual factor related to job performance using a single item. The trait or behaviour being assessed must also have been judged to bear some potential relation to the criterion measure. Also included are "structured scores" collected from referees considered unlikely to be familiar with existing or prior work performance of the participants.

4. Unstructured - this category included cases where overall subjective global ratings were provided by referees, cases where no standardised rating scale was used at all and studies reporting the measurement of traits and behaviours considered unlikely to be related to the criterion measure. (E.g., written comments subsequently rated for favourability.)

Additional potential moderator variables were also investigated. For example, assessments of personality vs assessments of work-related behaviour, concreteness of rating dimensions and referees' knowledge of prior or existing work behaviour.

Interrater reliability of coding for structure was assessed using the kappa statistic (Cohen, 1960). The present author and a second rater (an experienced military psychologist) independently coded the data. Agreement between the two raters was reasonably high (k = .79). Disagreements were resolved through discussion and arriving at a consensus regarding the appropriate categorisation. Reliability of coding for additional moderator variables was also assessed using kappa, or, where appropriate, the intraclass correlation. Reliability for these variables ranged from k = .58 to k = 1.00 with a mean of k = .82. The intraclass correlations ranged in value from R = .91 to R = .94. The levels of interrater agreement found in the present study are consistent with those reported by other authors (see Bullock & Svyantek, 1985; Wiesner & Cronshaw, 1988).

Overall, the post hoc analyses were unable to account for all of the variance observed in the study correlations but, as Table 3 shows, the pattern of results offers some support for a model of reference report validity based on structure. Highly structured reference reports were the most predictive of criterion outcomes (r = .21) and unstructured reports were the least predictive (r = .11). Structured (r = .17) and semi-structured (r = .16) fell in the middle ground with very little difference between these two categories. The estimated mean validity coefficients from the bare bones analysis are modest although there is a substantial improvement following corrections for the influence of statistical artifacts. The final row in Table 3 shows that the residual variance remaining after corrections for sampling error was non-significant for highly structured (χ^2 = 29.1, df = 22, p > .05) and structured reports (χ^2 = 36, df =

26, $p > .05$). However, the percentage of unexplained variance for semi-structured (42%, $\chi^2 = 100.3$, $df = 53$, $p < .001$) and unstructured reports (69%, $\chi^2 = 48.8$, $df = 14$, $p < .0005$) was considerable, and was highly significant in both cases. Therefore, additional moderator analyses were conducted.

Table 3
Meta-analysis of the predictive validity of reference reports moderated by structure

	Type of Report			
	Highly Structured	Structured	Semi-Structured	Unstructured
Total sample size[a]	4868 (23)	6093 (27)	25925 (54)	14801 (15)
Mean validity coefficient[b]	.21 (.41)	.17 (.35)	.16 (.34)	.11 (.24)
Sample weighted variance in corrected correlations	.019	.02	.013	.013
Sample weighted sampling error variance	.015	.015	.007	.004
% variance accounted for by sampling error	79	77	58	31
95% Credibility interval	.29 to .53	.23 to .47	.20 to .48	.05 to .43
χ^2	29.1	36	100.3*	48.8**

[a] The number of coefficients contributing to the total sample is given in parentheses. Outliers are excluded from this analysis.
[b] Means were corrected for restriction of range and criterion unreliability; corrected values are given in parentheses.
* $p < .001$.
** $p < .0005$.

Validity coefficients for each category of report were further sub-divided according to the type of criterion measure (performance versus training) used in the study. The results from this analysis are presented in Table 4.

Table 4
Meta-analysis of the predictive validity of reference reports moderated by structure and criterion type

	Type of Report			
	Highly Structured	Structured	Semi-Structured	Unstructured
Training Ratings				
Total sample size[a]	2540 (10)	954 (2)	9840 (18)	5673 (7)
Mean validity coefficient[b]	.23 (.41)	.16 (.28)	.14 (.26)	.07 (.13)
Sample weighted variance in corrected correlations	.003	.005	.011	.014
Sample weighted sampling error variance	.009	.005	.005	.004
% of variance accounted for by sampling error	100	100	45	29
95% credibility interval	-	-	.11 to .41	-.07 to .33
χ^2	3.3	-	39.6**	24.5*
Performance Ratings				
Total sample size[a]	2328 (13)	5139 (25)	16085 (36)	9128 (8)
Mean validity coefficient[b]	.17 (.35)	.17 (.35)	.18 (.37)	.14 (.30)
Sample weighted variance in corrected correlations	.028	.022	.01	.001
Sample weighted sampling error variance	.019	.017	.008	.003
% of variance accounted for by sampling error	68	77	82	100
95% credibility interval	.16 to .54	.21 to .49	.28 to .46	-
χ^2	19.1	32.4	45	2.7

[a] The number of coefficients contributing to the total sample is given in parentheses. Outliers are excluded from this analysis.
[b] Means were corrected for restriction of range and criterion unreliability; corrected values are given in parentheses.
* $p < .001$.
** $p < .005$.

The predicted relationship between reference report structure and validity was confirmed in the case of those studies utilising training criteria. Highly structured reports were the most valid followed in sequence by structured, semi-structured and unstructured reports. The mean validity for highly structured reports was more than three times greater than that for unstructured reports (.23 versus .07). Once again, as with the overall analysis, the difference in mean validity between structured and semi-structured reports was small (.16 versus .14). For highly structured and structured reports most of the variance in validity coefficients was accounted for by sampling error (see Table 4). For semi-structured and unstructured reference reports considerable variance remained after corrections for sampling error. The high levels of residual variance that were unexplained, coupled with the significant chi-square results, suggests that additional moderator variables may be present.

The pattern of results for those studies using performance criteria is not as straight forward. Table 4 shows that unstructured reports had the lowest mean validity, but that no clear differences are apparent between highly structured, structured and semi-structured reports. In fact, semi-structured reports performed marginally better than the others. The bulk of the variance in study correlations for highly structured, structured and semi-structured reference reports using performance criteria was found to be largely due to sampling error (see Table 4). Only structured reports were found to have less than 75% of the variance in validity coefficients accounted for by sampling

error. These findings, along with the non-significant results from the chi-square tests, make the presence of additional moderator variables unlikely.

Discussion

The results from the present meta-analysis offer some encouragement for those investigating reference reports. They demonstrate that under the right circumstances reference checks can achieve acceptable levels of validity for selection purposes. The data suggest that if practitioners and researchers are prepared to take the time to incorporate sound psychometric principles into the development and construction of reference reports then they can expect to see some returns on their investment. The corrected estimated mean validity for highly structured reports was 0.41, which is substantially greater than the mean validity reported by Hunter and Hunter (1984) for reference checks in their meta-analysis. Furthermore, with the exception of unstructured reports, the corrected mean validity for all types of reference check compares favourably with other alternative predictors that have been reviewed (see Hunter & Hunter, 1984; Reilly & Chao, 1982). Moreover, for the most part, lower bound credibility intervals do not include zero, suggesting that reference checks are able to provide non-zero validity coefficients. However, the magnitude of some of these coefficients may still be too small to be of any practical value.

The results from the present study show that definite gains in validity can be realised if structured reference reports are used in a systematic fashion with attention being paid to sound test development principles. Procedures such as forced-choice may yield particularly high dividends. The relative superiority of forced-choice over other formats appears to be explicable on at least two grounds. The first stems from the reason why they were designed and used in the first place, that is, to control for leniency errors and the resulting restriction of range that typically arises. If forced-choice reports are even only partially successful in that capacity, then it is obvious that the lower mean validity for other types of report could be at least partly due to this additional artifact. This explanation is consistent with findings reported in recent meta-analyses evaluating the interview (see Marchese & Muchinsky, 1993; Wright et al., 1989; Wiesner & Cronshaw, 1988). In particular, Wiesner and Cronshaw were able to show that the superiority of structured interviews over other forms was largely due to their improved psychometric properties, in their case, improved reliability. Unfortunately, the value of forced-choice rating forms for the control of lenient responses is difficult to ascertain as none of the researchers using this approach for the design of reference checks have specifically reported its success in reducing leniency effects. In fact, although leniency in ratings has been postulated as one of the most common and significant problems associated with the use of reference reports, there is no research available to date that has evaluated the accuracy of ratings provided by referees.

An alternative, but related, explanation for the superiority of forced-choice rating forms focuses on the methodology underlying the development of forced-choice reports and their resulting content. Smith and George (1992), in a comprehensive review of the personnel selection literature, emphasised the link between job content and the content of the selection method. In general, the greater the congruence between these two factors the greater the likelihood that predictive validity will be enhanced. Wiesner and Cronshaw (1988) in their meta-analysis provided some evidence for this conclusion by demonstrating that interview validity improved along with the job relatedness of interview questions. It seems plausible that researchers adopting a forced-choice approach are more likely to attend to actual job requirements and incorporate them in their questionnaire. The derivation of a forced-choice rating form encourages the researcher to produce job-related items and suitable distracters (see Guilford, 1954; Zavala, 1965), increasing the probability that the necessary links between the content of the job and that of the selection instrument will be achieved.

While the use of forced-choice scales show promise, they do have associated drawbacks. They are expensive and time consuming to develop in comparison to standard rating scales or unstructured forms. Furthermore, there can be resistance, or confusion, on the part of some referees when completing such forms, especially if they contain negative items (e.g., Rhea et al., 1965). Taylor and Wherry (1951) recount similar reactions by army personnel using forced-choice rating forms for performance appraisal purposes. In fact, they report that reactions were so extreme that the forced-choice rating scheme

had to be abandoned and an alternative rating scheme implemented. If reactions such as these are common, then there is a need for psychometrically sound rating formats that are effective as forced-choice scales, but more acceptable to referees. Furthermore, if alternatives should prove to be cheaper and easier to design and implement than forced-choice rating forms, then these alternatives would be welcomed by employers.

Although the results from the present study are encouraging, and suggest reference reports can contribute in a meaningful way to the selection process, some caution is required in their interpretation. Given the limited number of studies available, the influence of second order sampling error could be substantial. Estimates of mean effect sizes and the amount of variance attributable to artifacts can be affected by restricted sampling, even in cases where all available studies have been included in the analysis (Hunter & Schmidt, 1990). Population means and standard deviations for reference checks may differ somewhat from those reported in the present study due to the limited number of coefficients on which the analyses were based. For this reason the present results must be considered tentative. Further validation studies should be conducted, which can in turn contribute to a more "complete" meta-analytic review to be implemented at some later date.

The majority of the studies used in the meta-analysis were completed more than 30 years ago. The most recent publication was 1982. Considerable time has elapsed since these studies were produced which raises questions about their relevance to contemporary organisations in the 1990s, and the durability

of the results. Furthermore, many of the coefficients that were contributed were from the studies carried out in the armed forces. This limitation may pose problems regarding external validity, especially if the results from the meta-analysis are to be generalised to private sector organisations. These points further underscore the necessity for additional validation studies. Until such studies are completed, and can be incorporated into a meta-analysis, doubts about the robustness of the current findings will continue.

There are other reasons for interpreting the results from the present meta-analysis cautiously. For example, only after the validity coefficients were fully corrected for the influence of statistical artifacts were creditable levels of validity achieved. The corrected results provide some idea of the true validity of reference reports, but whether or not this is achievable in applied settings remains moot. If one considers the data from just the bare-bones analysis, the mean validity of reference checks, with the exception of highly structured reports, is not particularly heartening. However, this is not surprising given the major practical difficulties facing those envisaging using reference checks for selection. Even when well designed objective rating forms are used, there are additional sources of error that can easily creep in to distort the data. In particular, one must take into account the motivation of the referee, the likely absence of any foundation for comparability in the knowledge base of referees, and the diverse methods by which the reference checks are administered.

The motivation of raters has long been thought to play a significant role in the occurrence of leniency effects in performance judgements (e.g., Banks &

Murphy, 1985; Bass, 1956; Hauenstein, 1992; Murphy & Cleveland, 1995). According to this perspective, leniency is due largely to a rendering bias on the part of the rater. Raters may, for a variety of reasons, be unwilling to report their judgements accurately. Reference checks appear to be particularly susceptible to the influence of a rendering bias. In cases where reference reports are solicited, applicants are generally free to choose who will evaluate them. As such, it is reasonable to surmise they will select as referees those individuals they believe will provide a positive assessment. In the face of such expectations, there can be strong pressures on referees to avoid negative statements about applicants which may prejudice their chances of success. This reluctance on the part of referees to provide negative ratings is largely responsible for the severe leniency effects and consequential restriction in range problems that plague research in this area. Freedom on the part of applicants to select their own referees leaves the motivations of those providing ratings open to question. Employers may be quite justified in treating information collected from referees with some measure of suspicion. However, these suspicions may be mitigated should the referee be known (and respected) by those evaluating the applicant, or alternatively, if some or all of the information about the applicant is negative. The provision of negative information may be particularly significant in the interpretation of reference reports given the usual expectation for positive evaluations.

The motivation of the referee is only one of the difficulties facing researchers. Unlike more established psychometric procedures, the way reference reports are utilised by organisations and other employers for selection purposes is

rarely standardised. Furthermore, because applicants are responsible for nominating their own referee there are no guarantees concerning the quality and consistency of the information that is provided. Applicants will differ in their access to referees, and referees will differ in their ability to evaluate applicants, the opportunities to make evaluations, and their ability to communicate the results of any assessments they have made. These extraneous factors add unwanted variability to the process, contribute to the poor reliability of reference information and are likely to limit its validity.

Rating form structure does not account for all of the variance in validity coefficients from studies investigating reference reports. This suggests that other factors, such as those discussed above, may also moderate the validity of reference reports. Additional research to clarify the influence of potential moderators is called for. The performance rating literature provides an appropriate framework to draw upon as a basis for any further studies[1]. The majority of published studies on reference reports have investigated issues of reliability and validity. This is not too surprising as the criterion-related validity model is prevalent in much of the research on employment predictors. However, while the nominal function of reference reports is the prediction of subsequent performance on the job, the process of completing such reports usually involves some kind of rating of performance. Effectively, referees are asked to judge an applicant and then to provide evaluations (usually ratings) based on those judgements. In theory, employers will then use those evaluations to assist in the selection process. The action of rendering

[1] I am grateful to the anonymous reviewer who pointed this out.

evaluations on the part of the referee is clearly critical. It is surprising, then, that very little research has focused on the rating behaviour of referees.

Research on ratings in the context of performance appraisals is plentiful, however. An extensive amount of research has been conducted providing a useful foundation for any investigation of ratings supplied by referees. For example, the measurement of rater accuracy and rater errors is an integral aspect of research into performance ratings which could be applied to studies investigating reference reports. In particular, the calculation of accuracy measures offers a suitable methodology for assessing leniency in ratings from referees. None of the published studies to date on reference reports have examined the accuracy of referees.

Studies on performance rating can also contribute in other ways. Reviews of the performance rating literature (Decotiis & Petit, 1978; Ilgen, Barnes-Farrell, & McKellin, 1993; Landy & Farr, 1980; Murphy & Cleveland, 1995) have documented numerous factors that influence performance evaluations. Many of these factors, such as purpose/motivation, bias, and rater affect also appear to be relevant to ratings in the context of reference reports, yet their impact remains to be established. The motivation of the referee has already been discussed. A related issue is the question of whether or not raters are influenced by the mere fact that they are completing ratings for a reference report. Are raters who are completing evaluations for a reference report likely to be more lenient than those completing them for some other purpose, such as a performance appraisal? The problem of bias in reference reports is also

of interest. Are referees influenced in any systematic way by characteristics of the ratee, such as gender or race? The question of rater affect also deserves attention. Do referees rate those they like more leniently than individuals they dislike or feel neutral toward? Finally, the matter of rating form structure also merits further exploration. It has been suggested that improvements in validity associated with increasing structure in reference reports, found in the present meta-analytic study, may be due to the control and reduction of leniency in ratings. Accuracy measures used in performance rating research offer a methodology for directly assessing this possibility. The accuracy of referees using different rating forms could be compared and the effectiveness of alternative formats for reducing leniency in ratings determined.

In summary, few studies have investigated variables that influence the rating behaviour of referees. If one of the goals of research is to understand why it is that reference reports are often inadequate, and if gains in the validity of reference reports are to be realised, then it is essential that the rendering of evaluations by referees be examined in more detail. The following chapter reviews research from the performance rating literature and identifies some critical aspects of the rating process pertinent to the rating behaviour of referees.

Chapter 4

Literature Review: Performance Rating

The literature on performance ratings is multifaceted, complex and voluminous, as is evident from recent review articles and books (e.g., Ilgen et al., 1993; Murphy & Cleveland, 1995). An exhaustive review of the rating literature would be a Herculean undertaking. Therefore, to reduce what is a very broad topic domain to more manageable levels, the present chapter focuses on those facets of the literature most germane to reference reports. The principal research streams that have developed out of the performance rating literature are briefly reviewed before considering in detail aspects of the literature relevant to ratings in the context of reference reports. Topics covered in detail include: the influence of purpose on ratings, the effects of ratee characteristics on ratings, the impact of scale format on ratings, and the influence of rater acquaintance and affect. Where possible, the significance of these aspects of performance rating will be considered in terms of rater errors, and, in particular, leniency.

Rater Training

Rater training has been used both as a vehicle to elucidate cognitive processes in performance evaluation (e.g., Sulsky & Day, 1992; Woehr, 1994) and as a method to improve the quality of ratings (e.g., Bernardin & Buckley, 1981;

Fay & Latham, 1982; Hedge & Kavanagh, 1988). The initial focus of research in this area was on the reduction of rater errors (e.g., Bernardin, 1978; Latham, Wexley, & Pursell, 1975). This approach, termed rater error training, was based on the premise that familiarising raters with common psychometric errors (e.g., halo, leniency, central tendency), and then encouraging raters to avoid such errors, would improve the quality of rating data and the effectiveness of performance evaluations.

The legitimacy of these assumptions, and the value of rater error training has been challenged by some researchers who have highlighted the paradoxical relationship that can exist between rater errors and accuracy (e.g., Bernardin & Pence, 1980; Cooper, 1981). This relationship is one in which the presence of rater errors (e.g., halo) is linked with improvements in rating accuracy, and the reduction of errors is associated with decrements in accuracy. Rater error training has also been dealt a blow by researchers who have questioned the usefulness of error measures as assessments of rating quality (Murphy & Balzer, 1989; Saal, Downey, & Lahey, 1980; Sulsky & Balzer, 1988). These factors have prompted investigators to explore alternative rater training methods.

Before considering the alternatives, it should be pointed out that not all researchers have agreed that rater error training should be abandoned. Latham (1986) has argued that demand characteristics, and training in inappropriate response sets, account for the adverse research findings on rater error training. Interestingly, Woehr and Huffcutt (1994) found evidence

to support Latham's contention in their quantitative review of the rater training literature. They discovered, much to their surprise and contrary to common wisdom, that rater error training actually resulted in improvements in rating accuracy.

One alternative to rater error training is what Smith (1986) and Woehr and Huffcutt (1994) have called performance dimension training. Performance dimension training is based on the premise that raters will be more accurate if they can recognise and use performance-relevant dimensions rather than relying on global judgements. Training of this type has typically involved alerting raters to the relevant performance dimensions by involving them in the development of the scale, or by reviewing the rating instrument before it is used for evaluations. Woehr and Huffcutt found performance dimension training to be moderately effective for the reduction of halo, but to evince only modest improvements in accuracy.

Frame-of-reference (FOR) training (Bernardin & Buckley, 1981) is similar to but rather more elaborate than performance dimension training. Like performance dimension training, it stresses the multidimensional nature of performance. Where it goes beyond performance dimension training is in its efforts to establish common evaluative standards between raters. FOR training usually involves (1) identifying and defining performance dimensions, (2) reviewing behavioural incidents illustrative of varying levels of performance on the different dimensions, (3) using the standards that have been established to practice rating and, (4) feedback on the accuracy of ratings. As

Athey and McIntyre (1987) observe, the outcome of such training is the standardisation of raters' perceptions of performance. Woehr and Huffcutt (1994) report that FOR training was the most effective training method they reviewed for increasing rater accuracy.

The final training strategy to be reviewed is behavioural observation training. It has its antecedents squarely in the cognitive domain (see Lord, 1985), and emphasises the importance of accurately observing ratee behaviour. It incorporates techniques that stress the value of observing and recording behavioural information (e.g., diary records). The assumption underlying this type of training is that improvements in the recall of ratee behaviour will, in turn, improve rating quality. This approach has usually deemphasised judgemental or classificatory accuracy in favour of measures of recall and recognition. This has raised some questions regarding what the appropriate criteria in appraisal research should be (see Murphy, 1991; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Padgett & Ilgen, 1989). In the end, one must agree with Murphy (1991) that the criteria adopted should reflect the purpose for which ratings are collected. If feedback of information to ratees is of prime concern, then observational accuracy may be deemed the most relevant criteria. If, on the other hand, judgemental outcomes are important, then classificatory measures will be most appropriate. Woehr and Huffcutt (1994) analysed the impact of behaviour observation training for both types of criteria. They found that training enhanced observational accuracy, and reported that it also improved rating accuracy.

Cognitive Processes in Performance Rating

Much of the early research into performance ratings focussed on instrumentation and the psychometric properties of rating scales (Landy & Farr, 1980). In an effort to broaden the scope and focus of performance evaluation research, Landy and Farr called for a moratorium on studies investigating rating format, advocating instead that more attention be paid to process issues. Other researchers at this stage were also beginning to emphasise the importance of process factors in the appraisal process (e.g., Decotiis & Petit, 1978; Feldman, 1981). This proved to be the stimulus for the development of a extensive body of research inquiring into cognitive issues in performance evaluation.

Broadly speaking, cognitive approaches have addressed themselves to three major aspects of information processing during the rating process. Ilgen et al. (1993) identify these as: (1) the acquisition of information about ratees, (2) the organisation and encoding of information in memory, and (3) the retrieval and integration of information along with the act of rating itself. Under the rubric of those three areas, researchers have studied a variety of diverse phenomena such as: the effect of initial impressions (Balzer, 1986), accessibility of performance prototypes (Kinicki, Hom, Trost, & Wade, 1995), primacy and recency effects (Steiner & Rain, 1989), the biasing effect of behavioural anchors (Murphy & Constans, 1987) and much more besides (see Ilgen et al. for a comprehensive review). Ilgen and his colleagues note that research into rater cognitions has advanced theoretical understanding of social cognition

and how it operates in the judgement process. The research has also contributed in a more pragmatic way by providing information pertinent to the practical implementation of appraisal systems in organisations. Recently, however, there have been some misgivings about what is seen as a preoccupation in the literature with the cognitive aspects of appraisal. It is felt by some authors that social, situational and interpersonal influences in the appraisal process have been neglected (Dipboye, 1985; Ilgen & Favero, 1985; Murphy & Cleveland, 1995; Schneider, 1991).

Rating Scale Format

In a seminal review of the performance rating literature, Landy and Farr (1980) noted that the simple graphic rating scale was one of the earliest type of rating form to be developed. They also pointed out that research literature examining the logic and development of graphic scales was extremely sparse. Historically, little attention had been paid to the vehicle by which raters' evaluations were collected. However, the advent of alternative rating formats initiated a spate of research in which it was usual for the new methods to be compared to the traditional graphic system.

A variety of different rating formats have been developed over the years, spurred on for the most part by efforts to improve the quality of rating data. The proper design of rating forms was thought to remove ambiguity from the rating process and to help raters record more accurate and valid evaluations

of others' performance. The goal of designing better instrumentation to eliminate as much subjectivity as possible from ratings has been an alluring one. Behaviourally anchored rating scales (BARS; Smith & Kendall, 1963), mixed standard scales (MSS; Blanz & Ghiselli, 1972), behavioural observation scales (BOS; Latham & Wexley, 1977), forced-choice scales, and latterly, distributional rating scales (Kane, 1983) are all examples of alternative formats that have been developed. Unfortunately, despite considerable initial promise, reviews of the rating literature have consistently failed to demonstrate any clear superiority for any one type of rating instrument (e.g., Kingstrom & Bass, 1981; Landy & Farr, 1980: Schwab, Heneman, & Decotiis, 1975). The lack of progress eventually led many reviewers to call for a change of focus along with suggestions that more attention should be paid to process issues in rating (Decotiis & Petit, 1978; Ilgen & Favero, 1985; Landy & Farr, 1980). Researchers have responded to these pleas and, consequently, over the last decade there have been relatively few studies comparing alternative rating formats.

Fay and Latham (1982) compared behavioural expectation scales, behavioural observation scales, and trait-based scales for their resistance to certain rater errors. No differences were apparent between the behaviourally based scales but both were more resistant to contrast and first impression effects than the trait-based scale. Participants also rated the behavioural observation scale significantly better in terms of practicality. Gomez-Mejia (1988) reports contrary results in his comparison of BARS and global rating scales. His study is of particular interest as it included an Australasian sample. Various

properties of the rating scales were compared including halo, test-retest reliability, incremental utility, rating dispersion, and criterion-related validity. BARS were not superior to the global scales on any of the dependent measures. In fact, the global scales showed more resistance to halo effects and slightly greater criterion-related validity. Gomez-Mejia concluded by questioning the value of BARS, especially in light of the time required for their development and the associated costs. Dickinson and Glebocki (1990) compared BARS and four different versions of the MSS-type scale. They found the MSS formats did not differ with regards to leniency or halo effects, and showed superior convergent and discriminant validity compared to BARS.

Jako and Murphy (1990) investigated the effects of judgement decomposition and compared the accuracy and interrater agreement of a distributional rating scheme and Likert-type graphic rating scales. They found that distributional judgements did not improve agreement or accuracy beyond that provided by direct evaluative judgements using graphic rating scales. Decomposition (requiring simpler and less complex judgements), on the other hand, yielded more reliable and accurate ratings for both rating approaches. Steiner, Rain, and Smalley (1993) compared a distributional rating scheme to ratings collected using BOS-type scales. No differences directly attributable to rating format were found for reliability, although mean ratings from the distributional scales were significantly lower than those from the BOS. No conclusions about leniency or severity in ratings associated with either of the formats could be reached since the analysis involved mean differences rather than comparisons with true scores. However, Steiner et al. were able to

conclude that the distributional rating format appeared sensitive to variability in performance.

Härtel (1993) investigated the influence of field dependence and field independence on rating accuracy and affective reactions associated with different rating formats. Field dependence (FD) refers to an individual's cognitive dependence on the external organisation of information. Individuals characterised as field independent (FI) have the ability to impose organisation on information independent of the form in which it is perceived (Härtel, 1993). She hypothesised that the rating scale format would moderate the rating accuracy of FDs. As she predicted, FIs were more accurate than FDs when ratings were collected using a variety of holistic scales. No significant differences in accuracy for FDs and FIs were present when a decomposed rating format (greater structure) was used.

Fox, Caspy, and Reisler (1994) looked at the effect of cautionary instructions, the inclusion of irrelevant performance dimensions, and the impact of positively toned, asymmetrical scales on leniency, halo, and the validity of ratings for self appraisals. Scale format was found to exert a considerable impact on ratings. The psychometric properties of ratings from asymmetrical scales were much improved over those from standard graphic scales. Rating distributions were enhanced, mean ratings were closer to the scale mid-point, and convergent validity was increased. While there were obvious differences in ratings associated with the different scale formats, there were also theoretical problems in their interpretation. As Fox et al. point out,

participants using the asymmetrical scales rated themselves numerically almost one half scale unit lower than those using the standard scales. However, if the semantic meaning of the scale labels is considered, then those using the positively toned, unbalanced scales rated themselves almost one half scale unit higher than those using the standard scales.

Summary

In the last decade and a half there has been little research comparing different rating formats. For the most part, the research that has been conducted provides scant evidence for the superiority of any one particular format. Nevertheless, promising avenues of investigation have been identified in some studies. For example, it would be premature at this stage to dismiss the potential of distributional rating schemes. There is evidence that distributional ratings are sensitive to variability in performance (Steiner et al., 1993) but questions do remain regarding their psychometric properties (Jako & Murphy, 1990). Interesting data pertinent to the issue of rating formats have been produced by Härtel (1993). Her study indicated that rating format interacted with other elements in the rating situation to affect rating outcomes. The multifactorial nature of the rating process has been stressed by other authors (e.g., Murphy & Cleveland, 1995) with some going as far as specifically highlighting the potential for scale formats to influence raters and ratings (e.g., Kingstrom & Mainstone, 1985; Ostroff, 1993).

Reviewers who have considered forced-choice rating forms agree that they are effective in reducing leniency in ratings, although some questions remain regarding validity (Landy & Farr, 1980; Zavala, 1965). The meta-analytic review of referee reports carried out as part of the present research supports the efficacy of forced-choice rating forms. Results indicated that forced-choice forms were superior to most others. Unfortunately, as mentioned previously, forced-choice rating forms have a number of drawbacks. They are difficult, time consuming, and costly to develop. There can also be considerable resistance to their use on the part of raters (e.g., Rhea et al., 1965; Taylor & Wherry, 1951). Furthermore, because forced-choice forms provide overall scores, the information tends to be less diagnostic, and, therefore, not as suitable for identifying particular strengths and weaknesses.

The study by Fox et al. (1994) highlighted an alternative to forced-choice forms that may be just as effective in reducing leniency effects. The use of positively toned, asymmetrical scales to combat leniency in ratings was originally suggested by Guilford (1954), but has received little attention by researchers. The encouraging results documented by Fox et al. suggests that further research may be long overdue. More specifically, it remains to be established if the results reported by Fox et al. for self-assessments can generalise to ratings by others. Moreover, the Fox et al. study was limited by its reliance on mean differences in ratings for the assessment of leniency effects. Differences between groups attributed to scale formats may actually reflect true differences in performance. Although the random allocation of

participants to conditions means this potential confound is unlikely, it remains a possibility.

Interpretation of the results was also hampered by the lack of true scores for comparisons. Without true scores it is difficult to say how accurate ratings are when associated with a particular format. Fox et al. cannot say with assurance if ratings from the standard scales were more lenient than those from the asymmetric scales. For example, it is feasible, given the characteristics of the high performing participants in their study, that ratings from the unbalanced scales were actually too severe. There is simply no way of determining if leniency in ratings was present, or quantifying the amount. All that can be said is that there were differences between the two groups. Further research employing standard rating stimuli with true scores that allow relevant accuracy measures to be calculated would be a major step forward.

Rater Acquaintance and Affect

Research on the influence of rater acquaintance and affect has mostly stemmed from investigations of performance appraisal ratings (e.g., Tsui & Barry, 1986), and latterly, studies on leadership, specifically the leader-member exchange model (e.g., Wayne & Ferris, 1990). The notion that rater acquaintance and liking may cause bias, or in other ways influence the quality of ratings, is certainly not new (e.g., Ferguson, 1949; Freeberg, 1969;

Knight, 1923). However, until recently, it has received relatively little attention on the part of researchers. Renewed interest in the influence of rater affect coincides with calls for more research on the social and situational context of evaluation and judgement (see Judge & Ferris, 1993; Murphy & Cleveland, 1995).

Early studies in this area tended to focus predominantly on rater acquaintance. Researchers assumed that rating quality would be enhanced with escalating degrees of rater-ratee acquaintance. The rationale for this is that raters with increased acquaintance would be more likely to have representative samples of behaviour on which to evaluate performance (Kingstrom & Mainstone, 1985). However, initial investigations into the effects of acquaintance provided mixed results. For example, both Knight (1923) and Ferguson (1949) found leniency in ratings of individual traits to be associated with increasing degrees of acquaintance on the part of their raters. Surprisingly, Ferguson also reports improvements in the validity of supervisors' ratings associated with increased familiarity with their subordinate. What was not considered in these early studies was the nature of the relationship that existed between rater and ratee. One of the first to examine the impact of this variable was Freeberg (1969). He conducted a laboratory experiment in which he manipulated the extent and nature of rater-ratee contact. He found improved validity for ratings completed by raters with task-relevant acquaintance compared to those with irrelevant acquaintance. Kingstrom and Mainstone distinguished between task and personal acquaintance in their investigation of rater-ratee acquaintance and

rater bias among sales supervisors. Personal acquaintance was assumed to include an emotional component, that is, positive or negative affect. Their study is one of the first published reports to address the issue of rater affect directly. They found, using BARS-type scales, that both personal and task acquaintance were related to favourability of ratings, with personal acquaintance most influential. However, for the overall measure of performance and for promotions, differences in the favourability of ratings were attributable to actual differences in sales productivity. They suggested that the influence of acquaintance could actually depend upon the rating format adopted.

Further evidence regarding the importance of feelings in ratings was provided in a field study conducted by Tsui and Barry (1986). They explored the contribution of interpersonal relationships between raters and ratees to the quality of rating data. The affective attitude of their raters was operationally defined as the summated score on an affective relationship scale. The scale consisted of three items measuring the extent to which a rater admires, respects, and likes a ratee. Quality of ratings was evaluated in the light of several rating errors. They found that raters with positive affect were the most lenient, and those with negative affect were the most severe. Furthermore, the influence of affect on leniency was the same regardless of the source of ratings. Superiors, peers, and subordinates were all prone to rating inflation where positive interpersonal relationships existed, and to severity in ratings when negative affect was present. Raters with positive or negative affect were also more likely to produce ratings with halo. Finally, range restriction in

ratings was found to be greatest for raters with neutral affect. Tsui and Barry suggested that this could mean that removing feelings from the rating process may actually result in poor differentiation of performance levels among ratees.

Using written vignettes and student raters, Cardy and Dobbins (1986) conducted a laboratory experiment in which levels of liking were manipulated, and then evaluated for their impact on the accuracy of performance ratings. They found that accuracy of performance ratings decreased under conditions where performance levels and affect both varied relative to conditions where affect was constant and performance varied. They interpreted their findings as indicating that liking is an integral dimension, that is, one that is difficult to isolate from performance dimensions.

Following a hiatus, a spate of contemporary research directly addressing, or relevant too, the issue of rater affect has been conducted. Duarte and associates (Duarte, Goodson, & Klich, 1993, 1994) and Wayne and colleagues (Wayne & Ferris, 1990; Wayne & Kacmar, 1991; Wayne & Liden, 1995) have examined interpersonal affect as part of their research into the leader-member exchange model. In two studies (one laboratory experiment and one field study) Wayne and Ferris confirmed, using LISREL, the influence of supervisor liking for the subordinate on performance ratings. Ratings of performance were also related to objective performance levels. Similar findings were reported in a subsequent study conducted by Wayne and Kacmar using student raters. Confederate subordinates engaging in impression management received higher performance ratings than subordinates who did

not use impression management. Although not directly assessed, the authors argued that the positive rating effects resulted from the influence of impression management tactics on the intervening variable of rater affect. Similar arguments stressing the role of affect were posited by Duarte et al. (1993; 1994) to explain rating inflation for employees with high quality exchange relationships that they found in their studies. The hypothesised influence of affect in the studies conducted by Duarte et al., and by Wayne and Kacmar, is purely speculative. The authors did not measure rater affect, and hence, its role as a mediating variable remains subject to testing.

Stronger conclusions can be drawn from a study by Judge and Ferris (1993). Using LISREL, they evaluated the adequacy of a model of the performance rating process incorporating several social and situational factors, one of which was supervisor liking for subordinates. As predicted, supervisor affect was related to performance ratings. Supervisors were more likely to give positive ratings to subordinates they liked than to subordinates they disliked. Moreover, by testing alternative models Judge and Ferris were able to rule out competing explanations. The results did not support the premise that it was high performers who were better liked, but instead confirmed the principle role of affect. Supervisors who liked their subordinates rated them more favourably. As a proviso, however, they noted that "true" measures of subordinate performance were not available. In a subsequent study, Ferris, Judge, Rowland, and Fitzgibbons (1994) used LISREL to test a similar model in which supervisor affect towards subordinates was at the hub. The results provided strong support for the hypothesised link between supervisor affect

and performance ratings. Moreover, they also found that the allocation of rewards and resources was related to supervisors' liking of subordinates.

Robbins and DeNisi (1994) considered the influence of affect on the cognitive processing of performance information. Using a complex laboratory design and student raters, they found evidence for the influence of affect on both the process and outcome of performance evaluations and judgements. Affect-consistent and affect-inconsistent performance were seen as more meaningful, and weighted more heavily, than affect-neutral performance. Although not directly tested in their study, the authors argued that perceptions of past performance may be just as influential, and may have greater practical utility than measures of affect.

Some researchers have questioned the pivotal role of affect in the evaluation process. Wayne and Liden (1995) utilised a longitudinal research design and structural equation modeling to test a model of the effects of impression management on performance ratings. Contrary to their expectations, they found no support for a relationship between a supervisor's level of positive affect directed toward a subordinate, and ratings of that subordinate's performance. They suggest that the link between affect and performance ratings found in previous studies could be an artifact of the cross-sectional designs that were used. More specifically, they argue that longitudinal designs can reduce common method variance in the assessment of liking and performance and that it was this factor which could have been responsible for a spurious association between the two variables. Although their results are

intriguing, they must be interpreted cautiously. As the authors note, there was some participant mortality over the course of the study, the effects of which are difficult to fully quantify. It is possible, for example, that participants who dropped out of the study were those with whom supervisors had developed negative affective relationships. Selective dropout of this type would dilute the influence of affect and make it difficult to detect any potential effects. The nature of the sample used in the study could also account for their anomalous findings. Participants were new employees whom supervisors were required to rate for likability only six weeks after joining the organisation. In such circumstances, assessments of likability may not be particularly stable. Supervisors who originally liked their subordinates could change their mind as they got to know them over the course of the six month employment relationship. Likewise, those who initially felt antagonistic or neutral towards their subordinate could also have changed their mind. It may take some time for positive or negative affect to develop and stabilise and assessments of affect prior to that point would be unreliable.

Further evidence questioning the link between affect and performance ratings has been provided by Borman, White, and Dorsey (1995). Using path analysis, they examined the contribution of a variety of interpersonal and contextual variables on performance ratings. To avoid problems of common method variance, different raters were used to rate performance and interpersonal variables. Some interpersonal variables were found to influence performance ratings. Ratee dependability had the strongest effect for both peers and supervisors. Ratee friendliness and likability appeared to have very

little direct effect on performance ratings. Although common method variance was reduced in this study by using different raters, this approach appears to pose some conceptual problems. Because ratings of friendliness were collected from one set of raters and evaluated for their influence on performance ratings collected from a different set of raters, the locus of affect is assumed to be the ratee. That is, positive affect (liking/friendliness) is implicitly regarded as a dispositional factor, or property of the ratee, which they bring to each rating situation. As a consequence, the Borman et al. study may have neglected what could be termed the relational aspect of affect (the positive or negative feelings a specific supervisor has toward a particular subordinate) in favour of a more general likability factor. Furthermore, it is also possible that the unique organisational setting and characteristics of the participants from whom they derived their data could have been a factor in their results. The armed forces are very structured and hierarchical organisations. Typically, "fraternisation" between ranks is discouraged. In such circumstances one could expect the influence of affect to be reduced. There may simply be fewer opportunities for interactions conducive to the development of positive or negative affect, especially between supervisors and subordinates. There is some evidence from the Borman et al. study to support this contention. Although measures of affect did not directly influence supervisors' ratings, measures of negative affect (obnoxiousness) were influential in the case of peer ratings. It should also be noted that the supervisors in their study had undergone extensive rater training and were using behaviourally based rating scales. Both of these factors could be influential in minimising the affective component in evaluations.

Summary

There is documented evidence from a variety of studies that raters tend to evaluate individuals they like more positively than those they dislike. Furthermore, several authors have argued that the influence of rater affect is tenacious (Cardy & Dobbins, 1986; Tsui & Barry, 1986), that is, it is inextricably linked to the process of evaluation. Although not all researchers would necessarily agree with this view (e.g., Borman et al., 1995), what is apparent is that it would be precipitous at this stage to rule out the influence of rater affect in the evaluative process. This is especially true for the present study which is concerned with leniency in ratings in the context of reference reports where affective factors might be expected to play a greater role.

Ratee Characteristics

The question of how characteristics of a ratee may influence ratings has been researched extensively. Factors such as age (Cleveland & Landy, 1981; Waldman & Avolio, 1986), education (Tsui & O'Reilly, 1989), sex (Maurer & Taylor, 1994; Pazy, 1986), and race (Pulakos, White, Oppler, & Borman, 1989; Turban & Jones, 1988) have all been examined for their potential to introduce bias into the rating process. Although a range of demographic variables have been explored, researchers addressing the issue of subgroup

bias in evaluations have been particularly interested in potential race and sex effects. These factors are amongst those most commonly identified in human rights and employment legislation as a basis for nondiscrimination (Human Rights Act, 1993; U. S. Equal Employment Opportunity Commission, U. S. Civil Service Commission, U. S. Department of Labor, & U. S. Department of Justice, 1978) and are the variables most relevant to the present study.

Race

Early studies investigating the effects of race on evaluations have reported mixed results. Some studies have documented a bias against blacks (e.g., Hamner, Kim, Baird, & Bigoness, 1974; Parsons & Liden, 1984), others a bias in favour of blacks (e.g., Schmitt & Lappin, 1980), and yet others no effects whatsoever (e.g., Schmidt & Johnson, 1973). A meta-analytic review of ratee race effects conducted by Kraiger and Ford (1985) shed more light on the issue and went some way toward reconciling the inconsistent findings recounted in the literature. They reported corrected mean correlations of .18 and -.22 between ratee race and ratings from white and black raters respectively. These results indicated the presence of a same-race bias in performance ratings. That is, there was a clear tendency for white raters and black raters to assign higher ratings to members of their own racial group. Kraiger and Ford go on to report that these results were moderated by the setting in which ratings were collected and the saliency of blacks in the sample. Race effects were found to be most likely in field settings (as opposed to laboratory-based studies) and when blacks comprised a small percentage of the workforce.

A potential limitation of the Kraiger and Ford (1985) review was their inability to disentangle the influences of performance and race, a point made by Oppler, Campbell, Pulakos, and Borman (1992) in their discussion of methodology. They identify various approaches that have been utilised for the assessment of subgroup bias in performance evaluations. The first of these, the total association approach, is characteristic of many field studies exploring bias in ratings. Typically, researchers attempt to estimate the amount of criterion variance accounted for by subgroup membership by comparing ratings given to white ratees to those given to black ratees. Unfortunately, such studies cannot distinguish between rater bias and true performance differences. Differences between subgroups can be attributed to real differences in performance, or to criterion contamination. Evidence conducive to the former explanation has been provided by Ford, Kraiger, and Schechtman (1986). They argue that the uniformity in effect sizes for both objective and subjective criteria found in their meta-analytic review implies that the race effects "found in subjective ratings cannot be solely attributed to rater bias" (p.334).

The second approach identified by Oppler et al. (1992) for the assessment of subgroup bias is the direct effects approach. According to Oppler et al., researchers using this approach have attempted to eliminate real differences in performance between members of different subgroups prior to any assessment of rater bias. Consequential differences in ratings are then more clearly attributable to rater bias. Laboratory studies (which are prevalent in

this category) have done this by controlling performance levels. The performance of ratees is usually held constant, or varied independently of race (e.g., Schmitt & Lappin, 1980). In field studies, true performance differences have been controlled by statistical methods. The influence of nonrating factors is removed from the ratings before comparisons are made between subgroups (e.g., Oppler et al., 1992; Pulakos et al., 1989). Interestingly, the results from studies employing these methodologies indicate that the effects of ratee race on performance evaluations may have been overstated. Kraiger and Ford (1985) estimated the corrected correlation between race and performance ratings for the laboratory studies they reviewed to be only .03. Pulakos and colleagues (Pulakos et al., 1989; Oppler et al., 1992) have found consistent rater and ratee race effects in the large army samples they have analysed. However, they estimate such effects account for less than 2% of the total criterion variance. There appears to be some consensus in the literature that race can influence performance evaluations, but that, in general, the magnitude of such effects is small.

Sex

The literature on gender-related bias in ratings has produced inconsistent results (Nieva & Gutek, 1980). Some studies have reported an evaluation bias in favour of females (e.g., Hamner et al., 1974; Mobley, 1982; Norton, Gustafson, & Foster, 1977). Other studies have reported no differences in ratings as a function of ratee sex (e.g., Cascio & Phillips, 1979; Pulakos & Wexley, 1983; Thompson & Thompson, 1985) while yet further studies have

reported a pro-male bias in ratings (Dipboye, Arvey, & Terpstra, 1977; Pazy, 1986).

These contradictory and ambiguous research findings make it difficult to draw any firm conclusions about effects in this area. However, some tentative statements do appear warranted. Firstly, unlike studies that have investigated the effects of race, those inquiring into gender typically have found no interactions. That is, there is very little evidence for any kind of same-sex rater-ratee bias (Izraeli & Izraeli, 1985; Mobley, 1982; Pulakos & Wexley, 1983). However, these results are complicated by recent findings from a study conducted by Tsui and O'Reilly (1989) who found the performance ratings of subordinates were affected by the degree of "relational demography" evident in superior-subordinate dyads. Increasing dissimilarity in six superior-subordinate demographic factors (of which sex was one) was associated with poorer performance ratings. However, it must be emphasised that the effect sizes reported in their study were minimal and that sex was only one of six factors which they considered. Overall, relational demography appeared to account for only a small proportion of the variance in ratings. These findings are consistent with the results of a study conducted by Pulakos et al. (1989) who found that ratings of army personnel were influenced by the sex of the ratee, but that the amount of variance accounted for by ratee sex was less than 2%. Although a pro-male bias in ratings was evident, in practical terms the effects appeared negligible.

It has been noted that many of the studies in which sex differences in ratings have been documented were conducted in laboratory settings, and that results from studies conducted in the field have been much less definitive (Dobbins, Cardy, & Truxillo, 1988; Maurer & Taylor, 1994; Pulakos et al., 1989). This has lead some researchers to abandon simple sex effects to consider other gender-related factors such as relational demography (Tsui & O'Reilly, 1989), sex-related stereotypes (Dobbins et al., 1988), gender-related occupational stereotypes (Bartol & Butterfield, 1976), and perceived masculinity and femininity (Maurer & Taylor, 1994).

Summary

Recent meta-analytic reviews (Kraiger & Ford, 1985; Ford et al., 1986) and studies using large samples (Pulakos et al., 1989; Oppler et al., 1992) have confirmed that race and sex do influence ratings, but report that the magnitude of such effects is small. In contrast, other studies have continued to emphasise the significant consequences of sex (Pazy, 1986) and race (Turban & Jones, 1988) for performance evaluations. Previous research has been criticised on methodological grounds (Dipboye, 1985; Oppler et al., 1992; Pazy, 1986) and for the paucity of field studies (Dobbins et al., 1988; Oppler et al., 1992). However, questions remain regarding the generalizability of more recent studies and in particular, the large scale analysis of army ratings conducted by Pulakos and associates (Oppler et al., 1992; Pulakos et al., 1989). The small effects for race and sex they report may be due to specific characteristics of the sample and the context in which ratings were

collected. More specifically, the raters in their studies had undergone extensive training and were provided with well constructed behaviourally anchored rating scales. Two further considerations are also relevant and would appear, at least prima facie, to offer potential explanations for the minimal effects that have been observed. The first is the sizeable representation of ethnic minorities in the U.S. army. If, as Kraiger and Ford suggest, racial saliency is a factor in biased ratings, then we would expect bias to be reduced in situations where ethnic minorities are prominent. Secondly, the United States has in place strict legal guidelines in relation to race and sex discrimination in employment situations. Employers risk stiff legal penalties should these be contravened in any way (U. S. Equal Employment Opportunity Commission et al., 1978). All of these factors would appear to mitigate against bias in ratings. In situations where these external constraints were not present, bias in ratings might easily arise.

Recent surveys of New Zealand managers (McGregor, Thomson, & Dewe, 1994) and women directors (Shilton, McGregor, & Tremaine, 1996) suggest that women are under-represented in management and senior management positions. Other authors (Chen, 1993) have highlighted anti-Asian feeling and discrimination in New Zealand. Such research alerts us to the ever present possibility of bias and discrimination in employment settings here in New Zealand. Therefore, before ruling out the prospect of sex and/or race effects in evaluations, further research in a New Zealand context is required.

Purpose of Rating

There have been relatively few empirical investigations of the effects of purpose on rating outcomes. An early study conducted by Taylor and Wherry (1951) found that army officers rated subordinates more leniently when evaluations were to be used for administrative purposes rather than for research purposes. However, a subsequent study which employed a similar instructional set, and also utilised military personnel as raters, found no significant differences between ratings completed for either purpose (Berkshire & Highland, 1953).

Likewise, inconsistent results have been reported from studies using student evaluations of teaching staff. Sharon and Bartlett (1969) found significantly greater leniency in ratings completed for evaluation purposes, or in cases when raters believed their evaluations would have to be justified to the appraisees, compared to ratings completed for research purposes only. The significant differences arising out of the purpose of the rating task were only distinguishable for evaluations completed using a graphic rating scale. Ratings completed using a forced-choice scale appeared resistant to any kind of leniency bias engendered by rating purpose. Similar results were reported in a study by Driscoll and Goodwin (1979). Student raters were found to be more lenient when evaluations were provided for administrative purposes (decisions regarding promotion, tenure, or salary) than when ratings were simply to be used for course improvement. Unfortunately, any conclusions regarding the effects of purpose in the Driscoll and Goodwin study are

tempered by the presence of higher order interactions with class size and level.

Centra (1976), on the other hand, found very little evidence that raters were affected by the stated purpose of the rating task. Ratings of the overall effectiveness of university teachers showed no significant differences for evaluations collected under the pretext that they would be used for administrative reasons as opposed to ratings collected supposedly for course improvement reasons. Similarly, Meier and Feldhusen (1979) found no differences in ratings of university teaching staff gathered for administrative as opposed to diagnostic purposes, and concluded that students can rate instructors consistently regardless of the eventual disposition of the ratings. Likewise, Borresen (1967) found no difference in self ratings, or ratings of others by peers, when the stated purpose of a rating task was varied.

Zedeck and Cascio (1982) examined the effect of purpose and training on performance ratings of supermarket checkout operators. They presented student raters with written vignettes describing job-related behaviours on five critical performance dimensions. Purpose was manipulated by written instructions leading students to believe that ratings were for one of three reasons: a merit pay increase, recommending development, or retention of a probationary employee. Raters who were evaluating for a merit pay increase provided ratings with less discrimination than raters assessing for development or retention. Ratings for a merit pay increase also differed from those for other purposes with respects to "global differential accuracy." A

greater proportion of the variance in ratings for merit pay increases appeared to be accounted for by true variation in performance on the critical job dimensions than was the case for ratings completed for other purposes. The authors concluded that training was not as influential as the stated purpose of the rating task, but that training for specific evaluative purposes could lead to more accurate evaluations because raters' cognitive strategies appeared to vary as a function of purpose.

Evidence supportive of Zedeck and Cascio's (1982) view has been supplied in a series of cognitively oriented studies that have utilised more traditional accuracy measures. Pulakos (1986) compared rater accuracy for different rating tasks following training. Student raters were allocated to one of three training conditions (control, evaluative, or observational) and to one of two rating task conditions (evaluative or observational). She found that accuracy improved when the rating task and training conditions were congruent. Raters who received training consistent with the type of rating task were more accurate than those who received no training, or training that was incongruent. Murphy, Balzer, Kellam, and Armstrong (1984) examined the effect of purpose on observational and evaluative accuracy. Student participants completed ratings of instructors for research purposes (research on scale formats) or decision purposes (decisions about teaching assistantships). They were required to evaluate teacher performance on eight dimensions and assess the frequency with which certain important behaviours occurred. No significant differences in observational or evaluative accuracy were apparent as a direct function of purpose. There was some

evidence, however, that purpose influenced the information processing strategies of the raters. In a similar study Murphy, Philbin, and Adams (1989) tested the accuracy and recognition memory of student raters who had viewed videotaped lectures. Purpose of the rating task was manipulated by instructing raters before viewing that they were to learn the content of the lecture (observational condition), or alternatively, that they were to evaluate the lecturer's performance (evaluation condition). Participants were further required to complete their ratings either immediately or after varying delays of up to seven days. When there was less than three days delay before supplying ratings, raters were generally more accurate when the evaluatory nature of the task was explicit. When the delay was greater than three days, raters were more accurate when the evaluatory nature of the task was incidental. Murphy et al. (1989) interpreted their results as evidence that the purpose of observation can affect both the encoding and retrieval of information. In two separate experiments, Williams, DeNisi, Blencoe, and Cafferty (1985) explored the effect of appraisal purpose and outcome on information acquisition and utilisation strategies adopted by raters. They found that raters were sensitive to distinctiveness and consistency information contained in performance vignettes, but that there were few differences in the utilisation of information by raters evaluating for different outcomes or purposes. A main effect for rating purpose was apparent, with raters evaluating for purposes of remedial training referrals rating higher than those for promotions or merit pay increases. Unfortunately, this result is difficult to interpret as the stimulus materials used by raters in the remedial

training condition were different from those used by raters in the other conditions.

Dobbins and associates (Dobbins, Cardy, & Truxillo, 1986; Dobbins et al., 1988) have extended research on the purpose of appraisal to consider its interaction with ratee sex and the role of gender stereotypes. In an initial study (Dobbins et al., 1986) they found that ratings were biased against women when they were completed for administrative purposes as opposed to feedback or research purposes. In a subsequent study (Dobbins et al., 1988), they also reported a bias against female ratees. Raters with traditional stereotypes of women evaluated female ratees less accurately, but only when ratings were for administrative purposes. Ratings for research purposes showed no such systematic bias.

McIntyre, Smith, and Hassett (1984) investigated the effects of rating purpose and types of rater training on various indices of rater accuracy. Participants completed ratings for one of three purposes: feedback for course improvement, research into rater accuracy, or hiring of graduate teaching assistants. Clear effects were established for rater training but there was only tentative support for the effects of purpose. Participants in the research condition were found to be more severe in their ratings than participants in either of the other conditions. Interestingly enough, raters across all of the conditions were found to be severe in their ratings. Conclusions were tentative because heterogeneity in the variance of ratings had prompted the application of a conservative analytic strategy. To offset the consequential

loss in power, the authors elected to adopt a non-conventional alpha level of .10. The effects of purpose were significant at the revised alpha level but the authors were cautious about drawing any definite conclusions regarding the influence of purpose.

In a recent study, Ostroff (1993) examined the influence of raters' perceptions of the purpose of appraisal, norms about the rating process, perceived fairness of the appraisal system, and job satisfaction on rating outcomes. She surveyed 64 managers who were responsible for evaluating a total of 340 employees. Each manager responded to a questionnaire designed to measure what they believed the purpose of appraisals were, their reactions to the appraisal system, norms about honesty in evaluations, and job satisfaction. She found that managers who believed the purpose of evaluation to be administrative were more likely to be lenient in their ratings than those who highlighted other purposes such as feedback, goal identification, or documentation. Contrary to expectations, those who endorsed norms about honest rating behaviour were also more lenient in their ratings.

The interesting feature of this study was its focus on perceptions. The explicit, or stipulated, purpose of the rating task was not varied, but Ostroff (1993) found that raters' perceptions of the appraisal process differed, and that these perceptions systematically affected their rating behaviour. The strength of her conclusions are moderated somewhat by several factors. "True" performance levels of the ratees were unknown, hence, differences in managers' ratings could actually reflect real differences in employee

performance. Furthermore, the effect sizes reported were small, although, as she notes, more variance in ratings may be accounted for in circumstances where the explicit purpose for rating also varied.

Finally, the reliability of some of the scales that were used were unknown, or in the case of the normative measures, on the low side. Nevertheless, the study is important and could go some way toward explaining the inconsistent findings regarding the effects of purpose on rating outcomes. Clearly, individual differences in perceptions of rating purpose and norms about rating behaviour could dilute the impact of manipulations of the explicit purpose of rating.

Rater Motivation

Research on the effects of purpose on rating outcomes is often taken as indirect evidence of the influence of rater motivation (Murphy & Cleveland, 1995; Salvemini, Reilly, & Smither, 1993). Raters are assumed to be differentially motivated to supply accurate ratings as a function of the rating task. The importance of rater motivation has been stressed by several authors who have presented models focussing on the rating process (e.g., DeCotiis & Petit, 1978; Mohrman & Lawler, 1983; Murphy & Cleveland, 1995). Motivation is also central to the distinction between evaluations and judgements that has been highlighted in the rating literature. According to Mohrman and Lawler, a particular evaluation, or rating, is not necessarily equivalent to a rater's judgement of performance. Judgements are private and

internal evaluations. They "exist inside the head" of the rater. Ratings, on the other hand, represent public evaluations. They can be scrutinised by others and are what appraisers are prepared "to put on paper." Murphy and Cleveland make a similar point in their discussion of the goal-directed nature of appraisals. They suggest that raters are able to evaluate accurately, but on occasions, choose not to do so. Deficiencies in ratings are more often related to the appraiser's willingness to provide accurate ratings, rather than their ability to do so. In their discussion they present a simple model of the motivational forces affecting rating behaviour. The motivation to rate accurately is seen as a joint function of the rewards available for accurate rating, and the probability of receiving those rewards. The motivation to distort ratings is a function of the negative consequences associated with accurate ratings, and the probability of experiencing those consequences. The final outcome, the rating behaviour that is exhibited, depends on the relative strength of these motivational forces.

There is strong evidence in the literature to support the integral role of motivation in the rating process, and the fact that raters may consciously distort their evaluations. The research on rating purpose provides indirect evidence of motivational effects. Murphy and Cleveland (1995) have suggested that ratings collected for research purposes usually correspond more closely to judgements (show greater accuracy) because there are few adverse consequences for raters or ratees. On the other hand, for appraisals conducted in organisational settings there are seldom rewards for accurate rating (Napier & Latham, 1986), and, furthermore, raters are frequently

required to furnish appraisal information to satisfy multiple, and often conflicting, organisational objectives (Cleveland, Murphy, & Williams, 1989). In addition, Longenecker and associates (Longenecker & Gioia, 1992; Longenecker, Sims, & Gioia, 1987) have pointed out that the politics of appraisal in many organisations can actually encourage raters to distort their ratings. Further evidence for the role of motivation is provided by a recent study that examined the influence of accountability in the rating process. Mero and Motowidlo (1995) have suggested that accountability can be viewed as a motivating force on raters. They found that raters who were accountable (believed that they would have to justify their performance ratings) were more accurate than raters who were not accountable. Accountable raters were also more likely to comply to situational pressures encouraging accuracy or leniency in ratings. Research by Ostroff (1993) is also pertinent to the issue of rater motivation. Her results indicated that factors with an implicit motivational aspect, namely rating norms and perceived purpose of rating, could influence rating outcomes. Finally, Salvemini et al. (1993) have directly addressed the question of rater motivation. They found that the accuracy of raters was significantly improved by the offer of incentives for accurate rating. Participants who were competing to earn monetary rewards for providing the most accurate ratings were more accurate than those who were offered no such rewards. Overall, the results from these studies emphasise how important motivation is in the rating process.

Summary

The purpose for which ratings are collected can affect the accuracy of ratings. In particular, ratings from studies in field settings which have made use of non-student raters and used ratees familiar to the appraiser appear to be most vulnerable to the influence of purpose (Murphy & Cleveland, 1995). Such studies typically report greater leniency in ratings that are to be used for administrative purposes. One problem with the interpretation of results from the field studies is the lack of experimental control. It is often difficult to say with any certainty if differences in ratings reflect the influence of purpose or if they represent real differences in performance. A related issue, the prevalence of leniency effects found in ratings from organisations, raises interesting possibilities. It is clear that lenient ratings could reflect bias in the rating process. Alternatively, one could argue that they are an accurate indication of true performance levels. Given the efforts on the part of most organisations to maximise the quality of their staff (e.g., through selection and training), it is not so surprising to find skewed rating distributions.

Further investigation of the influence of purpose on ratings is necessary. However, the effects of purpose need to be considered in a multivariate context. How does purpose interact with other variables to influence rating outcomes? The study conducted by Mero and Motowildo (1995) points to the complex nature of the rating process. Participants in their study were susceptible to situational cues endorsing particular rating strategies.

However, despite the obvious influence of the rating context, participants proved resistant to attempts to elicit favourable ratings for female ratees. Clearly, there may be some factors, such as characteristics of the ratee, that moderate the impact of motivational forces such as purpose. There is also evidence that the type of rating scale that is used can moderate the effects of purpose on rating outcomes. Several studies have shown that leniency associated with variations in the purpose of ratings is reduced or entirely eliminated when alternate rating formats are used (Fox et al., 1994; Sharon & Bartlett, 1969; Sharon, 1970). Other authors (e.g., Ostroff, 1993) have called for further research to consider the interaction between rating format and purpose. The effects of different types of purpose should also be explored in more detail. Researchers typically have compared ratings collected for administrative purposes with those collected for research purposes. Surveys (e.g., Cleveland et al., 1989) indicate that, in practice, ratings are rarely used for research purposes. Comparisons between purposes that are more likely to be encountered in organisational settings should be of potentially greater relevance to practitioners.

Rater Accuracy

The assessment of rater errors has been one of the most common approaches employed by researchers for the evaluation of rating data (Landy, 1989; Sulsky & Balzer, 1988). The presence of leniency, halo, or range restriction in ratings is assumed to reflect deficiencies in the rating instrument, or on the

part of the rater. Conversely, rating accuracy is usually inferred in the absence of any such errors. As such, the assessment of rater errors is really an indirect method of evaluating rater accuracy (Murphy & Cleveland, 1995). Although popular, reservations about the adequacy of rater error measures as assessments of rating quality have been expressed by a number of authors (e.g., Murphy & Balzer, 1989; Saal et al., 1980; Sulsky & Balzer, 1988). They have criticised rater error measures on the grounds that they are prescriptive, and rely on arbitrary assumptions about the true distribution of performance and what the real relationship is between various indices of performance. Furthermore, they also point out that there are different operational definitions of the various error measures and that conclusions regarding accuracy are often dependent on the particular measure adopted. In response to these criticisms, more attention has been paid to direct measures of rating accuracy. These measures are not without their own practical, theoretical, and methodological limitations (Sulsky & Balzer, 1988), but are widely accepted as improved indices of accuracy in comparison to error measures. Murphy and Balzer (1989) used meta-analytic techniques to evaluate the relationship between error measures and measures of rating accuracy. The average correlation between error measures and accuracy measures was found to be only .05. A subsequent regression analysis using error measures to predict accuracy scores found that the majority of significant regression weights were actually negative. The results suggested that the presence of rater errors were actually associated with improvements in accuracy scores. Murphy and Balzer concluded that the rater error measures they examined

were not valid indicators of accuracy and recommended that they should no longer be employed as such.

Chapter 4 has reviewed and discussed research from the performance rating literature. A number of factors that impact on ratings in performance appraisals have been identified, including rating purpose, ratee characteristics, scale format, and rater acquaintance and affect. The following chapter describes an experimental investigation specifically designed to explore leniency in reference report ratings, where several of the above factors are examined as potential moderator variables.

Chapter 5

Experimental Study

The primary objective of the current experimental study was to evaluate leniency in ratings for three different reference report forms. Additional factors that might moderate the accuracy of ratings were also investigated. These included the purpose of rating, sex and race of the ratee, and rater affect.

Results from the meta-analysis conducted as part of the present research suggested that structured rating forms, and in particular forced-choice scales, were an effective method for improving the predictive validity of reference reports. The superiority of forced-choice scales has been attributed to their ability to reduce leniency in ratings. Although forced-choice scales have met with some success, researchers have reported concomitant drawbacks associated with their use. They are difficult, time consuming and costly to construct and raters often seem to dislike using them. Furthermore, the information generated from forced-choice forms tends to be less diagnostic than that from other rating forms. Guilford (1954) suggested that positively toned, asymmetrical rating forms might also reduce leniency in ratings.

Results supporting Guilford's (1954) contention have been reported in a study by Fox et al. (1994). They found that the psychometric properties of self-report ratings collected using asymmetrical scales were improved compared to

standard rating forms. An important focus of the present study was to determine if improvements in ratings using asymmetrical scales could generalise from self-reports to evaluations of others. Three different versions of a rating scale incorporating differing levels of structure and varying in their specific efforts to combat leniency were compared. The three different scales were as follows: (1) Unstructured rating scale - important performance dimensions were identified and raters were asked to provide written comments under each evaluating the ratee; (2) Likert-type graphic rating scale - important performance dimensions were identified and raters were provided with simple Likert-type graphic rating scales to use for evaluations; (3) Positively toned, asymmetrical rating scales - important performance dimensions were identified and raters were provided with positively toned, asymmetrical scales to use for evaluations.

A review of the performance rating literature identified several variables that could influence leniency in ratings from reference reports. The purpose for which ratings are collected was one such factor. Raters appear motivated to provide more lenient ratings for some purposes as opposed to others. There is also some evidence that the sex and race of the ratee can influence rating outcomes. Finally, raters' liking of the ratee also appears to be an influential factor. Raters who like the individual they are evaluating appear more willing to render lenient ratings than raters who express less positive affect.

The assessment of leniency in ratings from reference reports typically has been based on error measures such as skewed distributions and elevated

mean ratings. The adequacy of error measures has been questioned by some researchers who have shown them to be poor indicators of accuracy in performance rating research, and have advocated that direct measures of rating accuracy be used instead (Murphy & Balzer, 1989; Saal et al., 1980; Sulsky & Balzer, 1988). Remarkably, however, the research literature on reference reports is devoid of studies that have directly examined the rating accuracy of referees. The present study uses direct measures of rating accuracy to compare the evaluations of raters in different experimental conditions. The calculation of both error measures and direct measures of rater accuracy allows for comparisons of the different indices of rater accuracy.

The following hypotheses are empirically based, derived on the basis of research evidence in the performance rating literature:

Hypothesis 1. Mean ratings are highest for the asymmetrical form and lowest for the Likert-type form.

Hypothesis 2. Raters using the unstructured rating form are the most lenient in their ratings, whereas raters using the asymmetrical form are the most severe in their ratings.

Hypothesis 3. Ratings for purposes of a reference report are more lenient than those for a performance appraisal.

Hypothesis 4. Raters who express greater liking for the ratees rate them more leniently than raters who express less liking.

Hypothesis 5. Ratings differ as a function of the sex and race of the ratee.

Chapter 6

Method

Procedural Overview

Participants read a written vignette describing the performance of a university lecturer on a number of critical teaching dimensions. They rated the lecturer for likability and completed evaluations of the lecturer's teaching ability using one of three alternative rating formats. One group of participants were lead to believe they were completing ratings as part of a performance appraisal and the other group that they were completing ratings as part of a referee report.

Participants

Participants were distance education students enroled in either an introductory management paper, an introductory organisational behaviour paper or a graduate organisation and management paper in the faculty of Business at a New Zealand university. Experimental materials and requests to participate in the research were sent out to 917 students (395 males and 522 females).

Usable responses were received from 288 respondents, a return rate of 31.4%. The age of those who returned questionnaires ranged from 18 to 57

years, with a mean of 34 years. Fifty-eight per cent of the final sample were female and 42% male.

Procedure

The impact of four variables was assessed in a 2 x 2 x 2 x 3 factorial design comprising: Sex of the stimulus person (Male vs Female), Race of the stimulus person (Caucasian vs Chinese), Purpose of rating (Performance appraisal vs Referee report) and Type of rating form (Unstructured vs Likert vs Asymmetrical). The manipulation of rating purpose was included as a control variable to allow for an evaluation of the effect on rating accuracy of completing assessments for a reference report. Each possible participant in the study was randomly assigned to one of the 24 conditions. A cover letter, which requested participation in the study, and experimental materials were then mailed out to the target sample. The cover letter explained the purpose of the research, set out the procedure to be followed and also included a consent form (see Appendix 1).

Participants were asked to complete the consent form, read the vignette and rate the stimulus person in the vignette for likability. They were instructed to place the consent form, vignette and likability ratings into the enclosed, prepaid envelope that was to be returned to the experimenter. They then completed their ratings of the stimulus person. It was emphasised that they should not refer back to the vignette when writing their comments or when

completing the ratings. After finishing the rating task, all materials were to be mailed back to the experimenter in the prepaid envelope provided.

Experimental Materials

Vignette

A written vignette describing the teaching behaviour of a university lecturer was used as the rating stimulus (see Appendix 2). The position of university lecturer has often been used in studies of performance rating (see Krzystofiak, Cardy, & Newman, 1988; Maurer & Alexander, 1991; Steiner et al., 1993; Woehr & Feldman, 1993). The vignette contained 24 critical incidents describing the lecturer's teaching behaviour. The greatest portion of the incidents were selected from a collection of behavioural statements and effectiveness ratings originally developed by Sauser, Evans, and Champion (1979, cited in Krzystofiak et al., 1988)[1]. They reported incidents related to five core teaching dimensions for university lecturers (Relationships with Students, Ability to Present, Interest in Material, Reasonable Workload, Fairness of Testing). For each dimension there were 50 behavioural incidents which had undergone a standard behavioural anchored rating scale (BARS) retranslation procedure.

---

[1] I am indebted to Dr. Krzystofiak who supplied copies of the behavioural statements and effectiveness ratings.

The remaining dimension (Ability to Organise and Plan) was included after a review of prior studies on the performance assessment of teacher behaviour. Aspects of organisation and planning have been included in many of these studies. For example, Woehr and Feldman (1993) included organisation of material as one of the five dimensions of teaching performance used in their study of information processing in performance appraisal judgements. Likewise, Maurer and Alexander (1991), and Smither, Reilly, and Buda (1988) used organisation of a lecture as one of the target performance dimensions in their studies of contrast effects in behavioural measurement. Both Hauenstein (1992) and Athey and McIntyre (1987) used organisation as one of four critical teaching dimensions in their studies of rating accuracy. Bannister and colleagues (Bannister, Kinicki, DeNisi, & Hom, 1987; Kinicki & Bannister, 1988) constructed BARS-type scales to measure teaching effectiveness and one of the important dimensions for which a scale was developed was organisation. Many other authors have also identified organisation as a dimension of teaching effectiveness (e.g., Champion, Green, & Sauser, 1988; Doverspike, Cellar, & Hajek, 1987; Krzystofiak et al., 1988; McIntyre et al., 1984; Murphy, Balzer, Lockhart, & Eisenman, 1985; Murphy & Constans, 1987; Murphy et al., 1982; Nathan & Lord, 1983). Therefore, it was felt that this was an important category that would have particular relevance for the distance education sample used in the present research.

The vignette comprised written statements, supposedly from students, highlighting the lecturer's ability in different aspects of the core teaching dimensions. Four incidents were selected to represent performance in each

dimension. For four of the six dimensions, there were two positive statements and two negative statements. Statements were selected on the basis that they resulted in a mean effectiveness rating as close to neutral as possible. Of the remaining dimensions, one contained four positive statements and the other four negative statements. It was felt that this arrangement best represented an average performance level maximising the opportunity for a rendering bias and leniency effect in the ratings. Additionally, it also allowed for a comparison of rating behaviour across dimensions representing different levels of absolute performance. Arrangement of the behavioural statements in the vignette was determined randomly.

Four alternative versions of the vignette were constructed. In each case the core teaching dimensions and behavioural statements remained constant while the description of the hypothetical lecturer was systematically varied. The descriptive information preceding the behavioural statements took the following form:

> Mary Goh is a lecturer at the university. She is Chinese and is thirty years old. She has been employed by the university for the last four years. Each year, the students' association and the university administration ask students to comment on the performance of the lecturers they come into contact with. The following excerpts are statements from these annual surveys describing the job performance of Mary Goh. You should take these

> excerpts as being representative of her everyday performance on the job.
>
> Please read the excerpts carefully. Once you have finished you will be asked some questions about them.

For half of the vignettes the lecturer was described as male and for the other half as female. The impact of race was examined by describing the lecturer as either Chinese or Pakeha (Caucasian).

Likability Ratings

Each participant rated the likability of the lecturer depicted in the vignette on a 5-point, Likert-type rating scale (see Appendix 2). Responses could range from "not at all likable" to "extremely likable." These ratings were completed immediately after each participant had read the vignette. Rater affect toward the ratee has been assessed in a similar way in other studies (e.g., Cardy & Dobbins, 1986; Duarte et al., 1993).

Demographic Questionnaire

After reading the vignette, all participants were requested to complete a brief, one-page questionnaire (see Appendix 3). The questionnaire asked for data regarding the participant's age, occupation, ethnicity, sex, university experience, and familiarity with performance appraisals and referee reports.

This information was collected to allow for the analysis of potential moderator variables that could impact on participants' ratings.

Rating Forms

Instructions for rating the hypothetical lecturer were systematically varied so that participants believed they were completing either a performance appraisal or a reference report. Each participant was asked to rate the lecturer described in the vignette using one of three alternative rating forms.

Unstructured Rating Form: Participants were free to comment on the teaching ability of the individual described in the vignette in any manner desired. The six core dimensions of teacher effectiveness were identified and used as prompts to help the respondents frame their remarks. Participants were also asked to provide an assessment of the lecturer's overall teaching ability (see Appendix 4).

Likert-type Rating Form: A Likert-type rating form was developed and used as another measure of rating response to the written vignette. The form consisted of the six core dimensions of teacher effectiveness and an overall measure of ability which respondents were required to rate on 10-point, Likert-type scales. Responses on each scale could range from very poor to very good. Participants were asked to complete their ratings using whole numbers only (see Appendix 5).

Asymmetrical Rating Form: This type of rating form was originally suggested by Guilford (1954) and was found by Fox et al. (1994) to reduce leniency in self-report ratings. The asymmetrical rating form is one in which only one unfavourable descriptor is provided. The rest of the scale comprises favourable descriptors of varying degree. According to Guilford, the premise on which the scale is based is the anticipation of a positive leniency error. Because raters are expected to be generous in their ratings, the scale is designed to allow for this and encourages a normal distribution of ratings around a higher rating category. This sort of scale may also be used in situations where the assessor's interest is directed toward only the more qualified respondents (e.g., awarding prestigious scholarships, selecting applicants for restricted programs, etc.). In the case of the present research, the scale categories adopted were *poor, okay, satisfactory, good, very good* and *excellent.* Participants were instructed to check the rating category that applied (see Appendix 6).

Leniency Scale

The Leniency Scale was developed by Schriesheim (1978; cited in Schriesheim, Kinicki, & Schriesheim, 1979) to measure and statistically control response bias on the part of raters. The scale is based on the premise that response bias is a stable characteristic of raters (similar to a personality trait) intrinsic to any rating situation. Bannister et al. (1987) go further and suggest that it provides an index of rater error that is "theoretically orthogonal" to ratee performance. If leniency effects are independent of ratee

behaviour, and can be quantified by the Leniency Scale, then the scale offers a convenient method for partialing out rater error variance without impacting on true score variance.

Unpublished studies cited by Schriesheim et al. (1979) provide evidence of internal consistency, test-retest reliability and construct validity for the scale. In subsequent studies, Schriesheim et al. report satisfactory reliability coefficients for the complete Leniency Scale and for a shortened version consisting of 11 items derived through factor analysis. They also found pervasive leniency effects in descriptions of leader behaviour when using both scales.

Bannister et al. (1987) found that partialing out leniency effects from the performance ratings of university instructors enhanced discriminant validity and slightly attenuated convergent validity relative to an uncontrolled analysis. This was in contrast to partialing out the influence of a measure of overall effectiveness which resulted in substantial decrements in both convergent and discriminant validities.

Highhouse (1992), on the other hand, has been critical of the Leniency Scale. He correctly points out that there has been very little research to support its use. In his own study he found that scores on the scale completed under one instructional set were unrelated to scores completed under a second instructional set.

Scores on the Leniency Scale seemed to depend on who was being described by the scale and were not independent of ratee behaviour (Highhouse, 1992). These results imply that the scale may not be measuring a leniency disposition at all. Clearly, additional research is warranted and the present investigation offered a ready opportunity for further scrutiny of the Leniency Scale.

Before the scale was sent out to participants, a minor adaptation to the wording of the items was completed. Personal pronouns were changed so that they no longer referred exclusively to males, and instead were consistent with the gender of the hypothetical ratee (see Appendix 7). Coefficient alpha for the Leniency Scale from the sample used in the present research was computed to be .79. The split-half reliability was determined using the unequal lengths Spearman-Brown coefficient (Norusis, 1992) and was found to be .77. The reliability of the scale is acceptable although the figures are a little attenuated compared to those reported by Schriesheim et al. (1979) who note that the scale has "internal consistency and test-retest reliabilities in excess of .85 in several samples" (p.12).

True Scores

True score estimates for performance in each dimension were generated by eight expert raters (staff members and graduate students familiar with the precepts of performance appraisal). Ratings were completed using the 10-

point, Likert-type graphic scale in the first instance. True score estimates were also generated using the asymmetrical rating form and were collected from the same expert raters 10 weeks after the initial ratings had been completed. This was considered a sufficient time period for each judge's ratings on the two scales to be independent.

The same procedure was followed for both rating forms. The expert judges were required to rate each of the behavioural incidents for each of the core teaching dimensions in turn. After rating all of the incidents for a particular dimension, they then provided an overall rating for that dimension. Finally, after completing ratings for all of the teaching dimensions, they were asked to provide an overall rating of performance.

Means were calculated for each dimension and used as the basis for true score estimates. Measures of interrater agreement (intraclass correlation coefficients: ICC) were computed (using ICC formula 2,1 from Shrout & Fleiss, 1979) and the judges' ratings examined for sources of disagreement. Overall agreement between raters for all items across all dimensions was .85 for the Likert-type form and .84 for the asymmetrical form. Statements from dimensions with poor agreement which had received divergent ratings were identified and then presented to the expert judges for re-rating. Ultimate ratings for those statements and for their associated dimension were eventually determined through discussion by all of the expert raters until a consensus was reached.

Dependent Variables

Accuracy Measures

A variety of accuracy and error measures were derived for the analysis of the present data. Drawing on Cronbach's (1955) formulations, overall rater accuracy (D^2) was broken down into component parts:

$$\textit{Elevation} \quad E^2 = (\bar{r}_{..} - \bar{t}_{..})^2 \tag{5}$$

$$\textit{Stereotype Accuracy} \quad SA^2 = \frac{1}{k}\sum_j \left[(\bar{r}_{.j} - \bar{r}_{..}) - (\bar{t}_{.j} - \bar{t}_{..})\right]^2 \tag{6}$$

where $\bar{r}_{.j}$ and $\bar{t}_{.j}$ = mean rating and mean true score for dimension j; $\bar{r}_{..}$ and $\bar{t}_{..}$ = mean rating and mean true score over all ratees and dimensions; k = number of rating dimensions.

Elevation refers to the accuracy of the average rating, over all raters and dimensions. Stereotype accuracy is the accuracy in discriminating among performance dimensions, averaging over ratees. Two additional components of Cronbach's D^2 measure, differential elevation and differential accuracy, could not be calculated from the data available due to the nature of the experimental design.

A modified form of the D^2 index, termed distance accuracy (DA), was also computed. McIntyre et al. (1984) operationally defined DA as follows:

$$DA_k = \frac{\sum_{j=1}^{r}\left[\frac{\left(\sum_{i=1}^{d}\left|T_{ij} - R_{ijk}\right|\right)}{d}\right]}{r} \quad (7)$$

where d is the number of items; r is the number of ratees; k is the subscript referring to the kth rater; R refers to the obtained rating; and T refers to the true score.

Leniency was operationally defined using another formula taken from McIntyre et al. (1984).

$$Leniency_k = \frac{\sum_{j=1}^{r}\left[\frac{\sum_{i=1}^{d}\left(T_{ij} - R_{ijk}\right)}{d}\right]}{r} \quad (8)$$

where d, r, k, R, and T are as defined as for the DA measure. The Leniency measure is similar to Cronbach's (1955) elevation component score, but as Sulsky and Balzer (1988) point out, is more descriptive in so far as it indicates the rater's tendency to be lenient or harsh, whereas the elevation score simply registers the presence of systematic differences between ratings and true scores. In fact, the distance accuracy and leniency formulas are

identical except for the fact that DA is based on the absolute difference between true scores and obtained rating, whereas the leniency score takes into account the sign, or direction, of any such difference.

Error Measures

In addition to the accuracy measures noted above, a variety of error measures used by Murphy and Balzer (1989) and originally reviewed in Saal et al. (1980), were calculated for each rater. In the case of the present study, estimates of leniency were of particular interest, although a measure of halo was also calculated. The following indices were used (a) MEAN: the absolute difference between the mean rating, over ratees and dimensions, and the scale midpoint (leniency); (b) LENMEAN: the difference between the mean rating, over ratees and dimensions, and the scale midpoint (leniency). [This is different from MEAN in that it takes into account the direction of any differences rather than simply using absolute values. It was thought to be a more appropriate indicator of leniency effects]; (c) SKEW: the skew of the distribution of ratings over ratees and dimensions (leniency). [Skewness was calculated using formulas provided by Downie and Heath (1965)]; and (d) SD: the standard deviation of ratings across dimensions (halo). [This operationalisation differs slightly from that adopted by Murphy and Balzer who used the variance in ratings rather than the standard deviation.]

Data Analysis

Rescaling of Ratings and Measures

To allow for comparisons between the different rating forms, participants' responses on the unstructured form were assigned numerical values. Three judges independently coded all statements using the relevant 10-point graphic rating scale for each dimension. Reliability of the judges' ratings was assessed using Pearson product moment correlations whilst agreement was measured using ICC formula 2,1 from Shrout and Fleiss (1979). Overall agreement among all raters was .98. Average correlation coefficients for the three judges ranged from .76 to .90 across the seven teaching dimensions (see Table 5).

Table 5
Mean reliability (average correlations) of judges' ratings for responses on the unstructured rating form

Teaching Dimension	Mean Correlation Coefficient
Organisation and planning	.79
Interest in material	.76
Relations with students	.85
Ability to deliver lectures	.83
Ability to assess work fairly	.87
Reasonableness of workload	.84
Overall teaching ability	.90

Prior to the analysis of variance (ANOVA) and examination of true scores, ratings were further transformed to maintain equivalence between forms and facilitate comparisons. Ratings from the unstructured and Likert-type rating forms were changed so that they used the same metric as the asymmetrical form. This was done by recoding all responses below the midpoint of the scale as "poor" and then allocating ratings of 6 through 10 to "Okay", "Satisfactory", "Good", "Very Good" and "Excellent", respectively. This transformation was also applied to the true scores derived from the expert judges. However, some of the error measures (particularly skewness) were very sensitive to changes in the unit of measurement. Hence, for the evaluation of error scores, ratings from the unstructured and Likert-type scales were left in their original metric.

To facilitate comparisons between different error and accuracy measures, the scores for these also had to be scaled consistently. For the accuracy measures, smaller scores denote greater accuracy. The leniency measure was recoded so that positive values indicate leniency in ratings and negative values represent severity in ratings. For the rater error measures, smaller values indicate the absence of that particular error. The only exception to this is the SD measure where smaller scores may be taken to suggest halo in the ratings. Correlations incorporating this measure were recoded so that positive values indicate a relationship where the presence of halo is associated with increases in other error measures, and decreases in accuracy.

Analyses

Data analysis was planned over several stages. Preliminary analyses focussed on simple descriptive statistics gleaned from the demographic questionnaire which provided information attesting to the qualifications and experience of the sample and allowed for conclusions regarding the generalisability of the results. It also furnished information concerning the use of referee reports and served as a basis for the identification of potential moderator variables that could impact on ratings. ANOVAs on ratings for each teaching dimension were carried out to evaluate the impact of the experimental manipulations.

The use of multiple analysis of variance (MANOVA) was considered but ultimately rejected given the exploratory nature of the research and the adverse impact such an analysis might have on statistical power. Ramsey (1982) has shown that the power of MANOVA diminishes as intercorrelations between the dependent variables increase. Cole, Maxwell, Arvey, and Salas (1994) have questioned this relationship and demonstrated that the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. However, they show that in several situations where there are moderate to strong positive intercorrelations between the dependent measures (as would be the case between different dimensions of teaching behaviour) then the statistical power of MANOVA decreases sharply. W. Hager (personal communication, June, 26, 1995) has argued on conceptual grounds against multivariate tests. He suggests that multiple dependent variables are often included in experiments for reasons of

economy rather than as a result of explicit hypotheses formulated by the researcher. Furthermore, he argues that while the interdependence between the dependent variables influences the multivariate test criteria, they do not invalidate univariate tests[2].

Following the ANOVA, various accuracy and error measures were calculated as a basis for comparisons between rating forms. The use of multiple measures allowed for the exploration of relationships between such measures. Finally, a regression analysis was conducted to identify demographic and dispositional influences on ratings, accuracy measures and error measures.

[2] A MANOVA was conducted and produced nearly identical results to the univariate tests.

Chapter 7

Results

Demographic Questionnaire

The sample was comprised of more women (58%) than men (42%). However, a closer examination of the data showed that the return rate for males and females was not significantly different ($\chi^2 = 0.19$, $df = 1$, $\rho > .05$). In terms of ethnicity, the great majority of respondents identified themselves as Pakeha (90.2%). The second largest ethnic grouping were those who identified themselves as Maori (4.9%). A small number of respondents could be classified as Asian (2.6%), Pacific Islander (1.1%) or Indian (1.1%). The disproportionate representation of Pakehas in the sample did not allow for further analysis of ethnicity. Fully 81.4% of the sample were in full-time, paid employment and nearly one third of the sample (27.9%) indicated by their current job title that they occupied a managerial position. Consistent with recent surveys (McGregor et al., 1994), women were significantly less likely to be employed in managerial positions. Of the male respondents, 34.7% described themselves as managers whereas only 23.4% of the women reported their occupation as manager ($\chi^2 = 4.48$, $df = 1$, $\rho < .05$).

Other items on the demographic questionnaire pertained to university experience, perceived value of reference reports and familiarity with

performance appraisals and reference reports. Tables 6 and 7 summarise the responses to the questions posed.

Table 6
Summary of responses to the yes/no items in the demographic questionnaire

Question	% Responding Affirmatively
Have you ever attended any university lectures?	73.6
Have you ever carried out a performance appraisal?	61.5
Have you ever been responsible for selecting new staff?	63.5
Would you ask for referee reports when selecting staff?	93.4
Would you ever hire someone who had a bad referee report?	50.5
Have you ever acted as a referee for someone?	72.2

Table 7
Descriptive statistics for questions from the demographic questionnaire

Item	Median	Standard Deviation
Number of university papers completed.	4.0	10.8
Number of performance appraisals completed.	3.0	38.6
Number of employment decisions involved in.	2.5	60.9
Number of occasions required to act as a referee.	2.0	11.7

It is clear from Tables 6 and 7 that those participating in the study were well equipped to do so. Approximately three-quarters of those responding had attended university lectures. Furthermore, the data showed that they had

completed a median of four university papers each. This information suggests that the participants, despite being distance education students, were in the main, well qualified to execute the rating task required of them. Their university experience implies that they are knowledgable about the relevant performance dimensions and familiar with many, if not most, of the critical incidents used in the vignettes.

The data also indicate that the sample were experienced at selecting new staff, conducting performance appraisals, and acting as referees. Over 60% of the respondents had conducted at least one performance appraisal and a similar number (63.5%) reported some involvement in the selection process for new employees. The median number of appraisals was found to be 3.0 and the median number of employment decisions in which the respondents were involved was 2.5. Standard deviations for the responses reported in Table 7 are substantial. This is a result of a skewed distribution and the presence of outlier values. A small proportion of respondents worked in the human resource area, and therefore had considerable experience in selection and appraisal. Likewise, because both graduate and undergraduate students were sampled there was sizeable variation in the number of university papers completed. Because of the skewed nature of the distributions, median values rather than means were judged the more appropriate measures of central tendency. It is noteworthy that if the analysis is restricted to that part of the sample who attested to experience in appraisals and selection, then the median values increase dramatically. For those who had conducted appraisals the median number increases from 3.0 to 10.0, and for those involved in selection decisions from 2.5 to 7.0. Much of the sample also had

referee experience (72.2%), with a preponderance of participants having acted in that capacity more than once.

Overall, it is apparent that participants in the present research were not only qualified as evaluators of instructor performance but were also familiar with the business environment. As a group they may be more representative of the "normal" population of raters in industry than the student participants typically utilised in other studies (e.g., Hartel, 1993; Jako & Murphy, 1990; Maurer & Alexander, 1991).

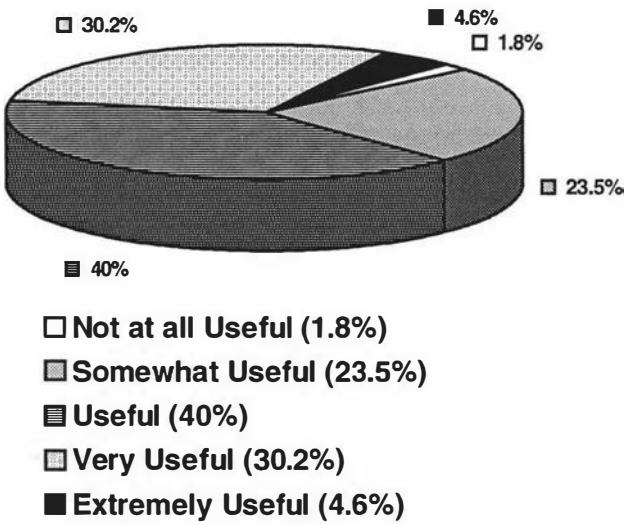


Figure 2
Rated usefulness of reference reports

Consonant with previous surveys (e.g., Muchinsky, 1979; Robertson and Makin, 1986), the great majority of the sample (93.4%) indicated that they

would ask for a reference report when selecting staff. The obvious willingness on the part of respondents to utilise reference reports during the selection process is mirrored by ratings of the usefulness of such reports. Figure 2 depicts the distribution of responses to the item asking participants about the usefulness of reference reports. Most of those questioned (74.8%) rated reference reports as "*useful*", "*very useful*" or "*extremely useful*". Only five of those responding (1.8%) considered reference reports of no use at all. Interestingly, almost half of those surveyed (49.5%) indicated that they would not hire an applicant with an unfavourable reference report. These results clearly affirm the importance attached to reference checks in the selection process.

Analysis of the participants' responses to the questionnaire when classified according to experience in the selection process, experience as a referee, and experience as a manager, highlighted some interesting features of the data. Respondents' experience in selecting new staff was categorised into three groups: no experience, some experience (those who had involvement in less than the median number of selection decisions for the sample, that is, less than seven) and very experienced (the remainder of the sample, that is, those with involvement in seven or more selection decisions). There were no significant differences between these groups in the likelihood of asking for a reference report ($\chi^2 = 0.29$, $df = 2$, $\rho > .05$), ratings of usefulness ($\chi^2 = 2.38$, $df = 8$, $\rho > .05$) or decision to employ an applicant with a bad report ($\chi^2 = 0.71$, $df = 2$, $\rho > .05$). Likewise, managers and non-managers were equally likely to request reference reports when selecting new staff ($\chi^2 = 0.53$, $df = 1$,

$\rho > .05$). Furthermore, both groups were equally inclined to employ an applicant even if that person received a bad report ($\chi^2 = 0.46$, $df = 1$, $\rho > .05$). As for ratings of the usefulness of reference reports, managers and non-managers were similar in their perceptions with no significant differences arising ($\chi^2 = 4.02$, $df = 4$, $\rho > .05$). Interestingly enough, however, none of the managers considered reference reports to be "*Not at all useful*". An anticipated result was that managers were significantly more likely to have acted as a referee. Almost all of managers surveyed (91.4%) had experience as a referee, whereas only 64.7% of non-managers had acted in that capacity ($\chi^2 = 20.57$, $df = 1$, $\rho < .00005$). Prior experience as a referee did not appear to impact directly on ratings of the usefulness of reference reports ($\chi^2 = 3.07$, $df = 4$, $\rho > .05$). There was a tendency for those with experience as a referee to indicate that they would be more likely to hire an individual with a bad report than those without such experience. However, this difference did not reach the conventional level of significance ($\chi^2 = 2.46$, $df = 1$, $\rho > .05$). Finally, experience as a referee did not seem to influence the decision to ask for a reference report when selecting new staff. Those with experience as a referee, and those without, indicated that they would be equally likely to require such a report ($\chi^2 = 0.80$, $df = 1$, $\rho > .05$). These results illustrate a remarkable degree of uniformity in perceptions amongst those surveyed. Experience in selection, or as a manager, or as a referee, did not seem to affect judgements about reference reports.

Not surprisingly, those who indicated that they would ask for reference reports when selecting new staff were more likely to rate them as significantly

more useful than those who indicated that they would not ask for such reports (χ^2 = 80.43, df = 4, $\rho < .00001$). Unexpectedly, 31.6% of those who indicated that they would not use reference reports still rated them as "*useful*". Those who believed referee reports to be either "*extremely useful*" or "*very useful*" indicated that they would be significantly less likely to hire an applicant if that applicant had received a bad report. Conversely, those who rated referee reports as "*not at all useful*" or only "*somewhat useful*" indicated that they would be significantly more likely to hire an applicant with a bad report (χ^2 = 27.78, df = 4, $\rho < .00005$). Evidently, for more than a third of those surveyed the decision to employ a job applicant would depend on the quality of the referee reports. This finding serves once again to reiterate the important role of reference checks in the selection process.

Tests of Hypotheses

ANOVA was used to test the following hypotheses: that ratings in the Performance Appraisal condition would be lower than those in the Referee Report condition, that ratings for the Male Stimulus condition would be different from those in the Female Stimulus condition, that ratings in the European Stimulus condition would be different from those in the Asian Stimulus condition, and that ratings would differ depending on the Form that was used. Ratings from the asymmetrical form were expected to be higher than those from forms 1 and 2. Ratings from the unstructured form were predicted to be higher than those from the Likert-type form. A familywise, modified Bonferroni test for planned comparisons, suggested by Keppel

(1991), was used to adjust alpha levels to correct for multiple tests. Critical alpha was set at $p < .01$ for the F-tests and $p < .004$ for the subsequent t-tests.

Table 8
Significant main effects for type of form from the overall ANOVAs calculated for each teaching dimension

Teaching Dimension	F value	df	p value
Organise & Plan	16.83	2, 264	$p < .001$
Interest in Material	14.77	2, 264	$p < .001$
Relations with Students	21.88	2, 263	$p < .001$
Delivery	32.63	2, 263	$p < .001$
Assess work Fairly	15.59	2, 263	$p < .001$
Workload	8.17	2, 264	$p < .001$
Overall Ability	17.78	2, 263	$p < .001$

A significant main effect for type of Form was found for each of the rating dimensions (see Table 8). Contrary to expectations, no other differences between groups were detected. Means and standard deviations for ratings from the different forms are presented in Table 9.

Results of the t-tests (one-tailed) comparing ratings are shown in Table 10. The results suggest a difference in rating behaviour dependent on the type of scale. Ratings from the asymmetrical form (Form 3) were significantly higher than those from the Likert-type scale (Form 2) for all seven dimensions of teacher behaviour that were evaluated. Similarly, ratings using the

asymmetrical form were significantly higher than those from the unstructured scale (Form 1), although in this case, for only four of the seven dimensions of teacher behaviour assessed.

Table 9
Means and standard deviations of ratings of teaching dimensions from three different rating forms

	Rating Form					
	Form 1 Unstructured		Form 2 Likert-type		Form 3 Asymmetrical	
Teaching Dimension	M	SD	M	SD	M	SD
Organise and Plan	3.37	1.07	2.64	1.35	3.59	1.28
Interest in Material	4.31	1.20	4.75	1.14	5.16	0.89
Relations with Students	1.81	1.03	1.19	0.52	1.98	1.03
Delivery	2.26	1.12	1.75	1.01	2.99	1.15
Assess work Fairly	2.84	1.27	2.71	1.48	3.65	1.25
Workload	2.53	1.10	2.48	1.48	3.15	1.41
Overall Ability	2.83	1.21	2.19	1.04	3.10	1.12

While significant differences were not apparent for all of the comparisons, it should be noted that mean ratings from the asymmetrical form were consistently higher, without exception, than those from Forms 1 and 2.

Significant differences between ratings were also apparent for comparisons between the unstructured scale and Likert-type scale. Use of the Likert-type scale resulted in lower ratings for the majority of the performance dimensions

(four of the seven). For three of the dimensions no significant differences in ratings were detected. Overall, the pattern of results suggest that raters using the asymmetrical form are the most lenient and those using the Likert-type rating scale the most strict.

Table 10
Results from t-test comparing mean ratings from three different forms

Teaching Dimension	Mean Scores		*t* value	*df*	Significance
	Form 1	Form 2			
Organise and Plan	3.37	2.64	4.05	179.42	< .001
Interest in Material	4.31	4.75	-2.54	180.00	*ns*
Relations with Students	1.81	1.19	4.80	105.14	< .001
Delivery	2.26	1.75	3.21	179.00	< .003
Assess work Fairly	2.84	2.71	0.63	179.00	*ns*
Workload	2.53	2.48	0.23	180.00	*ns*
Overall Ability	2.83	2.19	3.81	179.00	< .001
	Form 1	Form 3			
Organise and Plan	3.37	3.59	-1.25	182.00	*ns*
Interest in Material	4.31	5.16	-5.31	135.24	< .001
Relations with Students	1.81	1.98	-1.14	181.00	*ns*
Delivery	2.26	2.99	-4.29	181.00	< .001
Assess work Fairly	2.84	3.65	-4.28	181.00	< .001
Workload	2.53	3.15	-3.38	181.25	< .001
Overall Ability	2.83	3.10	-1.57	181.00	*ns*
	Form 2	Form 3			
Organise and Plan	2.64	3.59	-5.24	208.00	< .001
Interest in Material	4.75	5.16	-2.91	194.35	< .004
Relations with Students	1.19	1.98	-7.00	156.17	< .001
Delivery	1.75	2.99	-8.29	208.00	< .001
Assess work Fairly	2.71	3.65	-4.97	208.00	< .001
Workload	2.48	3.15	-3.37	208.00	< .001
Overall Ability	2.19	3.10	-6.10	208.00	< .001

Form 1 - Unstructured.
Form 2 - Likert-type.
Form 3 - Asymmetrical.
Alpha set at p < .004.

Accuracy Measures

Prior to the analysis it was hypothesised that there would be gains in accuracy for raters utilising structured rating forms. Changes in accuracy were also predicted for those completing ratings for performance appraisal purposes as opposed to reference checks. More specifically, ratings for performance appraisals were not expected to be as lenient as those for reference checks.

Table 11
Mean accuracy values as a function of purpose and rating form

Condition		Accuracy Measure [a]			
Purpose	Rating Form	*Elevation*	*Stereotype Accuracy*	*Distance Accuracy*	*Leniency*
Referee Report	Form 1 - Unstructured	0.86	1.04	1.22	0.69
	Form 2 - Likert-type	0.70	0.96	1.02	0.50
	Form 3 - Asymmetrical	0.61	1.06	1.02	0.07
Performance Appraisal	Form 1 - Unstructured	0.82	1.05	1.17	0.72
	Form 2 - Likert-type	0.62	0.94	0.94	0.28
	Form 3 - Asymmetrical	0.72	1.11	1.12	-0.12

a - Lower values denote greater accuracy.

Table 11 presents mean accuracy values for each rating form as a function of purpose of rating. Leniency scores show a notable and consistent improvement in accuracy for raters using different rating scales. For both referee reports and performance appraisals the asymmetrical form resulted in

the least amount of leniency in ratings. The unstructured rating form was the most lenient of the three rating scales. A particularly salient and intriguing feature of the data is that performance appraisal ratings from the asymmetrical form were, on average, actually stricter than the true scores supplied by the expert raters.

Table 12
Results of t-tests comparing mean accuracy values for ratings from three different forms

Accuracy Measure	Mean Scores [a]		*t* value	*df*	Significance
	Form 1	*Form 2*			
Elevation	0.84	0.66	2.13	178.00	*ns*
Stereotype Accuracy	1.04	0.95	2.00	178.00	*ns*
Distance Accuracy	1.19	0.98	3.27	178.00	< .008
Leniency	0.70	0.40	2.64	178.00	< .008
	Form 1	*Form 3*			
Elevation	0.84	0.67	2.10	132.28	*ns*
Stereotype Accuracy	1.04	1.09	-1.00	180.00	*ns*
Distance Accuracy	1.19	1.07	2.00	135.48	*ns*
Leniency	0.70	-0.04	6.30	180.00	< .008
	Form 2	*Form 3*			
Elevation	0.66	0.67	-0.19	198.12	*ns*
Stereotype Accuracy	0.95	1.09	-3.29	208.00	*ns*
Distance Accuracy	0.98	1.07	-1.72	198.03	*ns*
Leniency	0.40	-0.04	4.10	208.00	<.008

a - Lower values denote greater accuracy.
Form 1 - Unstructured.
Form 2 - Likert-type.
Form 3 - Asymmetrical.
Alpha set at p < .008.

Tests comparing ratings for referee reports and performance appraisals showed no significant differences for any of the accuracy measures used. Any differences in accuracy, should they exist, are clearly more attributable to the

type of rating form utilised. This interpretation is borne out by the data in Table 12, which presents the mean values for each accuracy measure for each form, and the results of t-tests (one-tailed) used to check for significant differences. Keppel's (1991) modified Bonferroni test was utilised once again to correct for multiple comparisons. The critical value of alpha for these comparisons was set at $p < .008$. Contrary to expectations, most of the accuracy measures showed no significant differences between rating forms. The exception to this was the leniency measure. Ratings completed using the asymmetrical form were significantly more accurate in terms of leniency than those from any of the other forms. Conversely, ratings completed using the unstructured form were significantly more lenient than those from any of the other forms. Another noteworthy feature of the data is the magnitude of most of the accuracy scores. For example, rater accuracy in distinguishing between performance dimensions (Stereotype accuracy) ranged from .95 to 1.09 across the three scales. In the context of a six-point rating scale, this represents considerable variation. In fact, all of the accuracy scores, with the exception of leniency values for the structured scales, are at least one-half and, more typically, one full scale unit away from true score values. Ideally, analysis of accuracy measures would be carried out using ratings in their original metric. Unfortunately, this introduces a potential confound in that the deviation measures from which the accuracy scores are derived are liable to be sensitive to the range of rating categories available. Replication of the analysis in the original metric lends support to this view (see Appendix 8). In the replicated analysis, differences between the forms are much more apparent, with structure clearly moderating all types of accuracy.

In conclusion, the data uphold Guilford's (1954) notion that the design of scales can mitigate against leniency effects. The asymmetrical form, designed according to Guilford's suggestions, resulted in significantly smaller leniency values than for either of the other two scales. However, when other accuracy measures are considered, differences between the scales largely disappear.

Error Measures

Table 13 presents the results of an analysis using error measures. Leniency was assessed by the measures MEAN, SKEW, and LENMEAN, and halo was estimated using SD. Differences between the leniency error values are apparent. The average LENMEAN and MEAN values show that the unstructured form yields the most lenient ratings, the Likert-type scale is somewhat less lenient, and the asymmetrical scale has the least amount of leniency in ratings. In fact, LENMEAN values for the asymmetrical form indicate that it actually encourages strictness (or severity) in ratings. On the other hand, the measure of skewness indicates that the asymmetrical form is the most lenient. Furthermore, the negative SKEW values for Forms 1 and 2 actually imply a tendency toward strictness in ratings when employing these scales. It appears that when using error measures to assess the quality of rating data, any conclusions regarding leniency in ratings will depend on the particular error measure adopted.

Because fewer rating categories were available for raters using the asymmetrical form, only Forms 1 and 2 allow for a direct comparison of halo in ratings. Table 13 shows that, regardless of purpose, the standard deviation of ratings is larger for the Likert-type form than for the unstructured form, indicating a greater spread in ratings, and consequently less halo.

Table 13
Mean error values for ratings from three different forms used for performance appraisal and reference report purposes

Condition		Error Measure			
Purpose	Rating Form	*SD*	*MEAN*	*SKEW*	*LENMEAN*
Referee Report	Form 1 - Unstructured	1.40	1.34	-0.09	0.97
	Form 2 - Likert-type	1.71	0.98	-0.17	0.78
	Form 3 - Asymmetrical	1.14	0.61	0.23	-0.02
Performance Appraisal	Form 1 - Unstructured	1.47	1.21	-0.26	1.06
	Form 2 - Likert-type	1.86	0.89	0.18	0.39
	Form 3 - Asymmetrical	1.19	0.73	0.19	-0.21
Purpose collapsed	Form 1 - Unstructured	1.44	1.27	-0.18	1.01
	Form 2 - Likert-type	1.78	0.94	-0.01	0.60
	Form 3 - Asymmetrical	1.17	0.68	0.21	-0.12

Comparisons Between Measures

While there exist different operational definitions of both rater error measures and measures of rating accuracy, the relationship between these measures is of some interest. Intercorrelations among the rater error measures are presented in Table 14. Correlations among measures representing "equivalent" leniency constructs have been highlighted in bold. Negative correlations characterise the relationship between SKEW and the other leniency measures. Only one positive coefficient was obtained for correlations between the leniency measures, and that was for MEAN and LENMEAN. The relationship between these two variables is sizeable, although by no means perfect. Quite clearly, any conclusions regarding the prevalence of leniency will depend upon the specific measure selected. Correlations between the leniency measures and the halo measure were generally quite modest and mostly negative.

Table 14
Correlations among rater error measures

Measure	*SD* [a]	*MEAN*	*SKEW*	*LENMEAN*
SD	--			
MEAN	.23	--		
SKEW	-.06	**-.08**	--	
LENMEAN	-.12	**.51**	**-.42**	--

a - *SD* measure is reverse coded.

Intercorrelations among the accuracy measures are shown in Table 15. Correlations among these measures are all positive, although there is

considerable variation in magnitude, ranging from weak to moderately strong. McIntyre et al.'s (1984) distance accuracy measure exhibits the strongest relationships, in this case with Cronbach's measures of elevation and stereotype accuracy (r = .75 and .66, respectively). Interestingly, the relationship between elevation and stereotype accuracy is weak (r = .12), bolstering the assumption that they assess different components of accuracy. McIntyre et al.'s leniency measure also appears to evaluate a unique component of accuracy. The strongest relationship it has is with elevation (r = .42).

Table 15
Correlations among rater accuracy measures

Measure	*Elevation*	*Stereotype accuracy*	*Distance accuracy*	*Leniency*
Elevation	--			
Stereotype accuracy	.12	--		
Distance accuracy	.75	.66	--	
Leniency	.42	.04	.32	--

Correlations between the accuracy and error measures are presented in Table 16. More than half of the coefficients (11 out of 16) are positive. The size of the correlations vary considerably, but most have a magnitude of less than .25. However, there are several notable exceptions to this. In particular, correlations between the accuracy measures and error measures MEAN and LENMEAN tend to be much more substantial. It is noteworthy that those correlations are all positive, probably due to the common derivation of these

measures from deviation scores. In fact, the almost perfect correlation (r = .99) between LENMEAN and Leniency, and between Elevation and MEAN (r = .98), suggests that they could be used interchangeably. Interestingly, the correlation between the two indices of halo (SD and stereotype accuracy) is negative despite the fact that both measures were scaled in the same direction. The remaining correlations are generally quite weak suggesting that the accuracy and error constructs are by no means equivalent.

Table 16
Correlations between accuracy and error measures

Measure	*SD* [a]	*MEAN*	*SKEW*	*LENMEAN*
Elevation	.24	.98	-.05	.43
Stereotype accuracy	-.21	.12	.02	.09
Distance accuracy	-.02	.74	-.07	.37
Leniency	.18	.49	-.41	.99

a - *SD* measure is reverse coded.

Regression Analysis of Error Scores

Using multiple regression, the four rater error measures were used to predict each of the four accuracy measures. The standardised regression coefficients and multiple R's resulting from this analysis are given in Table 17. It can be seen that two of the accuracy measures, namely *Elevation* and *Leniency*, were highly predictable from the combination of error measures used in the multiple regression. Adjusted R^2s of .97 and .99 were obtained for these two accuracy measures respectively. Almost all of the variance in these accuracy

measures was accounted for by knowledge of the relevant error measures. Furthermore, the positive, and highly significant standardised regression coefficients reported for MEAN and LENMEAN suggests that they may play a particularly important role in the prediction of the two accuracy measures. The close association between these particular error and accuracy measures is not surprising given the large correlations between them reported in Table 16.

Table 17
Results from a standard multiple regression analysis using four error measures to predict each of four accuracy scores

		Standardised Regression Coefficients			
Criterion	Adj R^2	*SD* [a]	*MEAN*	*SKEW*	*LENMEAN*
Elevation	.97	.02	1.02*	.00	-.09*
Stereotype Accuracy	.07	-.25*	.14*	.09	.09
Distance Accuracy	.58	-.20*	.80*	-.01	-.02
Leniency	.99	.06*	-.03*	.00	.99*

a - *SD* measure is reverse coded.
* $p < .05$.

The error measure for halo (*SD*) also contributed to the prediction of accuracy scores, but in such a way that the presence of halo errors was more typically associated with improvements in accuracy. Two of the three significant regression weights for this measure were negative. The implications of this are that high rater error scores (connoting the presence of errors) actually tends to indicate that raters were accurate rather than inaccurate in their rating. Scores for the final error measure (SKEW) appeared to be unrelated to any of the accuracy measures. None of the regression weights for this measure were significant.

Prediction of Leniency

The contribution of several personal and contextual elements of the rating situation to the prediction of leniency in ratings was explored. Rater affect was one of the important variables to be assessed. It was hypothesised that rater affect would influence ratings in such a way that those who expressed greater liking for the ratee would be more lenient in their evaluations. The contribution of a personal disposition towards leniency on the part of raters was assessed using the leniency scale devised by Schriesheim (Schriesheim et al., 1979). It was predicted that raters identified as characteristically lenient responders would in fact render more lenient ratings than those identified as not possessing a characteristically lenient response bias.

Several additional aspects of the rating situation believed to contribute to leniency in ratings were also examined. Prior analysis had shown that the format of the rating scale was closely associated with lenient responding. The same analysis indicated that rating purpose may also influence rating outcomes, although conventional significance was not attained. The present analysis allowed for a specific quantification of the effects of these factors and permitted a more detailed assessment of their relative contribution in combination with other variables. It might also be expected that experience in rating situations would affect rating outcomes. To explore this possibility, rater experience as a referee and with performance appraisals was included in the analysis.

A standard multiple regression was conducted using McIntyre et al.'s (1984) leniency value (Lenkmean) as the dependent variable, and rating form (Form), score on the leniency measure (Lentotal), likability of the ratee (Likability), purpose of rating (Purpose), experience with performance appraisals (Appraisal) and experience as a referee (Referee) as independent variables. The planned analyses allowed for the specific assessment of the criticality of each of these variables to the dependent measure.

Table 18
Correlations among variables included in the standard regression analysis for the prediction of leniency in ratings

Variable	Lenkmean[a]	Form	Lentotal	Likability	Purpose	Appraisal	Referee
Lenkmean	1.00						
Form	-.33*	1.00					
Lentotal	.36*	-.09	1.00				
Likability	.40*	-.07	.36*	1.00			
Purpose	-.14	.06	-.14	-.05	1.00		
Appraisal	-.11	.04	.07	.02	-.00	1.00	
Referee	-.02	-.05	.03	-.03	-.02	.57*	1.00

a - Smaller values denote greater accuracy.
* $p < .001$

Examination of residuals led to transformation of some variables to improve normality, linearity and homoscedasticity. The variables Appraisal and Referee were subject to a logarithmic transformation. Several transformations of the variable Likability were attempted to offset heteroscedasticity, none of which were particularly effective. As a consequence, the variable was left untransformed. One outlier case with a Mahalanobis distance greater than 25 was omitted from the analysis. Following suggestions from Tabachnick

and Fidell (1989), the regression coefficients and correlations between independent and dependent variables were examined for evidence of suppressor variables, but none were found. Table 18 presents the correlations between the variables.

Three variables show significant correlations with McIntyre et al.'s (1984) leniency measure. These were rating form (Form), leniency score (Lentotal) and ratings of likability (Likability). The significant negative correlation with Form suggests that as the rating scale changes from unstructured to Likert-type to asymmetrical, leniency in ratings decreases. This finding is consistent with analyses previously conducted. The significant positive correlations of the dependent variable with scores from Schriesheim's (Schriesheim et al., 1979) leniency scale (Lentotal) and measures of rater affect (Likability) indicate that as scores on these measures increase, so too does leniency in ratings. Of the correlations among the independent variables, only two are significant. The log of Appraisal and the log of Referee show a moderate to strong degree of association. This is not surprising given that managers and supervisors are much more likely to be approached with requests for reference checks. Performance appraisals are also more typically conducted by those in supervisory or managerial positions. Managers and supervisors with experience are simply more likely to have had opportunities to fulfil both functions. Ratings of Likability and scores on the leniency scale also show a significant positive correlation.

Results from the standard multiple regression are displayed in Table 19. Included are the standardised regression coefficients (β), the semipartial correlations (*sr* and sr^2), the partial correlations (*pr* and pr^2), R, R^2, and adjusted R^2 (adj R^2).

The multiple *R* for the regression analysis was .56, a highly significant value [$F(6, 239) = 18.61$, $p < .0001$]. Overall, 32% (30% adjusted) of the variance in McIntyre et al.'s (1984) leniency score was predictable from knowledge of the six variables included in the regression equation. Four of the six regression coefficients were found to be significantly different from zero.

Table 19
Results from a standard multiple regression using six personal and contextual variables to predict leniency in ratings

Independent Variable	*R*	R^2	Adj R^2	β	*sr*	sr^2	*pr*	pr^2
All variables	.56	.32	.30					
Form				-.27**	-.27	-.07	-.31	-.10
Lentotal				.22**	.20	.04	.24	.06
Likability				.31**	.29	.08	.33	.11
Purpose				-.08	-.08	-.01	-.10	-.01
Appraisal				-.15*	-.12	-.02	-.15	-.02
Referee				.06	.05	.00	.05	.00

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

The significant regression coefficient for Appraisal indicates that those with more experience in performance appraisal tend not to be as lenient in their ratings [$F(6,239) = 5.32$, $p < .05$]. The significant coefficients for Lentotal [$F(6,239) = 14.52$, $p < .0005$] and Likability [$F(6,239) = 28.99$, $p < .0001$] show that increases in scores on these variables are associated with more lenient

ratings. Finally, the significant regression coefficient for Form [F (6,239) = 25.38, $p < .0001$] shows that leniency in ratings is related to the type of rating form used to collect those ratings. Experience as a referee and purpose of the rating task did not contribute significantly to variability in the leniency of ratings. Although it did not reach the conventional level of significance, the regression coefficient for Purpose [F (6,239) = 2.23, $p < .14$] is in the expected direction. There is the suggestion that ratings completed for a reference report tended to be more lenient than those completed for a performance appraisal.

The squared semipartial correlation coefficients (sr^2) reveal the unique variance accounted for by each of the independent variables. Three of the variables (Form, Lentotal, and Likability) are able to account independently for 19% of the variability in the leniency measure. The unique contribution of the remaining variables amounts to only 3%. The six variables in combination contributed another 10% in shared variability.

In order to clarify further the influence of the leniency scale and rater affect, two hierarchical regression analyses were carried out. In the first, the contribution of the leniency scale was evaluated after controlling for the type of rating form. In the second, the contribution of the leniency scale was assessed after controlling for both the type of form and ratings of Likability. The results from both of these analyses are presented in Table 20. Included are R, R^2, adjusted R^2 , and change in R^2 (sr^2) as each variable was entered into the equation.

The initial step for both hierarchical analyses was the same, the entry of Form. The change in R^2 at this point was highly significant [R^2 = .11, F_{inc} (1, 244) = 28.85, p < .0001]. For the first analysis, the inclusion of scores from the leniency scale added significantly to the prediction of the dependent variable [R^2 = .21, F_{inc} (1,244) = 33.66, p < .0001]. The addition of ratings of Likability resulted in a further significant increment in prediction [R^2 = .30, F_{inc} (1,244) = 27.88, p < .0001], as did the inclusion of the log of Appraisal for step 4 [R^2 = .31, F_{inc} (1,244) = 4.95, p < .05].

Table 20
Results from the hierarchical multiple regression analysis for the prediction of leniency in ratings after controlling for the effects of rating form (Analysis 1) and for the effects of rating form and Likability (Analysis 2)

Independent Variable	*R*	R^2	Adj R^2	Chg R^2 (sr^2)
Analysis 1				
Form	.33	.11	.10	.11**
Lentotal	.46	.21	.21	.11**
Likability	.54	.30	.29	.08**
Appraisal	.56	.31	.30	.01*
Purpose	.56	.32	.30	.01
Referee	.56	.32	.30	.00
Analysis 2				
Form	.33	.11	.10	.11**
Likability	.50	.25	.25	.15**
Lentotal	.54	.30	.29	.04**
Appraisal	.56	.31	.30	.01*
Purpose	.56	.32	.30	.01
Referee	.56	.32	.30	.00

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

Similar results were found for the second analysis. In this case, step 2 required the inclusion of ratings of Likability. A significant improvement in prediction, over and above that for Form alone, was apparent [$R^2 = .25$, F_{inc} (1,244) = 47.66, $p < .0001$]. The addition of scores from the leniency scale at step 3 also resulted in significant improvements [$R^2 = .30$, F_{inc} (1,244) = 14.88, $p < .0001$]. For both analyses the addition of Purpose and Referee to the equations did not provide any significant improvement in R^2.

The pattern of results from the regression analyses suggests that Likability may be more influential than scores on the leniency scale for the prediction of leniency in ratings. However, the leniency scale continued to add to prediction even after controlling for the effects of both Form and ratings of Likability.

Chapter 8

Discussion

The experimental component of the present study investigated the relationship between raters' perceptions about the purpose of evaluation, rater affect, characteristics of the ratee, format of the rating instrument and rating outcomes. The use of positively toned, asymmetrical scales was found to reduce leniency in ratings compared to unstructured and Likert-type rating forms. Rater affect was also found to be related to leniency in ratings. Raters who expressed greater liking for the ratee showed a marked inclination to award more lenient evaluations.

Contrary to expectations, raters' perceptions about the purpose of evaluation did not exert a strong effect on rating outcomes. There was a tendency for those completing evaluations for reference report purposes to be a little more lenient in their ratings, but this effect was weak and accounted for very little of the variance in rating outcomes. Likewise, no systematic bias in ratings as a function of ratee characteristics was apparent. Significant ratee race and sex effects were not detected.

The Rating Instrument

The results from the present study show that raters are sensitive to variations in rating format. Unstructured rating forms that required raters to provide narrative comments produced the most lenient ratings. Ratings from Likert-type graphic scales showed significantly less leniency, and, at the same time, even greater improvements in accuracy were found for positively toned, asymmetrical rating forms. These findings are consistent with the results from the present meta-analysis, which showed improvements in predictive validity for structured reference reports, and, in particular, for structured reports which incorporated specific techniques to reduce leniency.

The current study corroborates results reported by Fox et al. (1994), and illustrates that gains in rating accuracy associated with the use of asymmetrical scales can generalise from self-assessments to evaluations of others. Fox et al. were unable to account fully for the abatement of leniency they found following the use of asymmetrical scales. The possibility of a self-protection motive on the part of raters was discussed, and although it could not be ruled out completely, was eventually dismissed as inconsistent with their data. The present study confirms that the changes in leniency reported by Fox et al. cannot be attributed solely to efforts by raters to avoid an overly positive self-presentation. If that were the only explanation for the gains in accuracy found by Fox et al., then one would not expect to see reductions in leniency for raters who are evaluating others. Improvements in rating accuracy for tasks other than self-assessments suggest that it is

characteristics of the rating form itself that are responsible for changes in leniency.

The realisation that raters are sensitive to variations in scale format is especially heartening for those seeking improvements in ratings from referees. The design of appropriate rating instruments is one of the few options available to organisations that are keen to influence the quality of rating data they receive from referees. This is because traditional methods for improving the quality of ratings are not easily applied in the case of reference reports. Training and the selection of competent raters are two common techniques that have been employed with some success to remedy rater errors. They are feasible options in many rating situations; however, they may not be practicable for referees. The reason for this is that referees are usually selected by the applicant, and would not necessarily be willing, or able, to participate in extensive rater training programs. Furthermore, even if some referees were willing to undergo training, it seems unlikely that organisations with an eye to the "bottom line" would count the money as well spent. Rater training programs are often expensive, and cost-benefits for the organisation are maximised when they can utilise their trained raters repeatedly. Training referees would not pay dividends for an organisation, except in the unusual circumstance where multiple applicants all nominated the same referee.

Results from the present study indicate that the introduction of positively toned, asymmetrical scales into reference reports may well be worthwhile for organisations seeking improvements in rater accuracy. Such scales are easy

to construct and simple to use. Moreover, the results show that reductions in leniency are possible without having to resort to expensive and time consuming alternatives, such as forced-choice rating forms. Potentially, positively toned, asymmetrical scales represent a very cost-effective solution to the problem of lenient ratings from referees.

Although the results from the present study are promising, they must be qualified at this stage. Following the introduction of asymmetrical scales, changes in leniency were not reflected by similar changes in other measures of accuracy. The ability of raters to discriminate among performance dimensions (stereotype accuracy) was unaffected by modifications of the rating instrument. Furthermore, because a number of Cronbach's (1955) accuracy measures (differential elevation and differential accuracy) were not utilised for the present study it is not known if they are influenced by changes in rating format. Further research incorporating other measures of accuracy is required. A related issue concerns the relationship between accuracy and criterion-related validity. It is unknown at present if gains in the accuracy of referees' ratings will translate into improvements in validity. Accurate ratings are obviously desirable, but may not guarantee validity. The relationship between different measures of rating accuracy and criterion-related validity remains open to question, and should be addressed in future research. Finally, further research is necessary to identify the mechanism by which rating scale format affects rating outcomes. Guilford (1954), who originally proposed the use of unbalanced rating scales to counteract lenient tendencies, offers no clear explanation for why such scales might be

efficacious. The question of why is it that positively toned, asymmetrical rating forms produce ratings with less leniency than other scale formats is yet to be resolved.

Rater Affect

Results from the regression analysis provide ample evidence for the influence of rater affect. Positive rater affect was associated with higher ratings, and negative affect was associated with lower ratings. Raters who expressed liking for the ratee were more lenient in their ratings compared to raters who expressed neutral or antagonistic feelings toward the ratee. These results are all the more compelling given the meagre information available to raters. Descriptions of the hypothetical ratee were brief, and contained mostly factual statements related to job performance. Yet, even under such artificial conditions, raters formed affective responses, and these responses were systematically associated with rating outcomes. It is reasonable to surmise that the influence of rater affect might be even stronger in situations where the ratee was known to the rater, and the relationship between the two was well established.

The results from the present investigation lend support to those from previous studies which have documented similar effects for rater affect (e.g., Judge & Ferris, 1993; Tsui & Barry, 1986). It should not be surprising that feelings influence our judgements, for according to Zajonc (1980), affective reactions

are "the major currency in which social intercourse is transacted" (p. 153). He submits that there are very few of our perceptions and thoughts that do not contain a significant affective component. Although an idealised view of the rating process might see it as cold and objective, it is moot as to whether or not one can avoid the "primary, basic, inescapable, and irrevocable nature of affective reactions" (Zajonc, 1980). Evidence from the present study suggests that the appraisal of others would not be exempt.

Although it can be argued, as Zajonc (1980) has done, that affective reactions are ubiquitous, the question of how they influence rater evaluations, and the full extent of their influence remains unclear. Studies have shown that other factors in the rating situation may moderate the power of affect. For example, Salvemini et al. (1993) found that the accuracy of raters was significantly improved by the provision of contingent rewards for accurate rating. Mero and Motowildo (1995) have shown that raters who are held accountable for their evaluations are more accurate than those who are not accountable. These studies suggest that the motivational context in which ratings are collected is important, and may assist raters to separate objective judgements from subjective feelings. This notion is consistent with assumptions underlying the goal-directed model of rating behaviour, outlined by Murphy and Cleveland (1995). Forces that comprise the motivational context, such as incentives or accountability, represent rewards or punishments that influence rating outcomes.

Emphasis on the motivational context of the rating process has interesting implications in relation to reference reports. Referees do not normally have any special obligation to the organisation requesting an evaluation of an applicant. They are not accountable to that organisation for their ratings, and there are unlikely to be any personal repercussions if they are inaccurate. In other words, there are very few extraneous forces encouraging accuracy on the part of referees, and because of this, reference reports may be particularly vulnerable to the influence of rater affect. In the absence of external pressures, raters may allow themselves to be swayed by personal feelings. Cognitive dissonance theory (Festinger, 1954) offers one explanation for why raters may be susceptible to affective reactions in such circumstances. Cognitive dissonance theory suggests that when our attitudes, values or beliefs are in conflict with our actions we experience psychological tension or dissonance. This tension is unpleasant; therefore, we seek to minimise or eliminate it. Research has shown (Festinger & Carlsmith, 1959) that dissonance is reduced if there are obvious explanations to account for the discrepancy between our beliefs and actions. If such explanations are not available, then we may have to change our behaviour or our attitudes in order to reduce tension. For referees, dissonance may arise when they like the individual they are evaluating, but must rate them poorly if they are to be accurate. Likewise, dissonance may be present when referees who dislike the ratee are obliged to provide positive ratings if they are to be accurate. In the absence of strong justifications for accurate ratings, referees may modify their evaluations so that they are consistent with personal feelings, and thereby reduce tension. The possibility that the influence of rater affect on rating

outcomes is mediated by cognitive dissonance is purely speculative, but merits further attention.

The variable of rater affect may account for disparate findings from studies that have evaluated the Leniency Scale. Results from the present study indicate that scores on the Leniency Scale are associated with lenient responding by raters. Participants who scored highly on the Leniency Scale tended to be much more lenient in their ratings. These findings are consistent with results reported by Schriesheim et al. (1979) and Bannister et al. (1987). They used the Leniency Scale to control statistically for response bias in leader descriptions (Schriesheim et al.) and performance evaluations (Bannister et al.). However, Highhouse (1992) has questioned the validity of the scale. He found that the scale had poor test-retest reliability, and that responses depended upon the specific target of the scale. Highhouse argued that if the scale measured a lenient disposition, responses should show a greater degree of consistency, and should be relatively independent of the rating target.

A parsimonious explanation for these conflicting findings is that the Leniency Scale represents an alternative measure of rater affect. The Leniency Scale consists of items derived from an instrument that assessed raters' tendencies to respond in a socially desirable manner. Respondents are obliged to answer true or false to items such as "No matter who he's talking to, he's always a good listener" and "She has never shown intense dislike for anyone" (see Appendix 7). It is plain that the items comprising the Leniency Scale are not

affectively neutral. Social desirability, by its very nature, must contain an affective component. Furthermore, the way in which the items are worded requires participants to make judgements about a particular stimulus person. In the present study this was the hypothetical lecturer described in the vignette. In other studies, judgements have been based on "real life" managers, supervisors, and university instructors. In all of these studies, participants have associated their ratings with particular individuals. This represents an emotional reaction to a specific stimulus, what Murphy and Cleveland (1995) have called directed affect. The end result is that the Leniency Scale may be doing nothing more than assessing raters' affective reactions toward specific ratees. Therefore, the gains in accuracy reported by Schriesheim et al. (1979) and by Bannister et al. (1978) may actually be due to the statistical control of rater affect. The significant correlation between the Leniency Scale and the measure of affect used in the current study lends further support to such a notion. The Leniency Scale as a measure of affect would also explain the results obtained by Highhouse (1992). If rater affect is being measured by the Leniency Scale, then one would expect scores to vary depending upon the selection of the target ratee.

A straightforward test of this proposition would be to ask respondents to complete the scale, rating individuals towards whom they have strong positive or negative affective reactions. If the Leniency Scale is assessing rater affect, then one would expect to see systematic changes in scores associated with the nature, and strength, of any affective reactions. A more rigorous test would be to use experimental manipulations to engender affective reactions in

participants directed towards a particular stimulus. Baseline leniency scores could be collected and then be examined for changes following the manipulation of rater affect. If leniency scores for the same rating stimulus change along with affective reactions, then this would be convincing evidence that the scale does not measure a stable predisposition on the part of raters to respond leniently, but is instead a measure of rater affect.

The regression analyses conducted as part of the present study showed that scores on the Leniency Scale were related to rating outcomes. However, interpretation of the changes in R^2 from the hierarchical analyses suggests that there is considerable overlap in the variance accounted for by the Leniency Scale and ratings of Likability. This is consistent with the view of the Leniency Scale as an alternative measure of rater affect. However, the Leniency Scale continued to contribute to the prediction of leniency, even after controlling for raters' liking of the ratee. This suggests that the scores from the Leniency Scale are not identical to ratings of Likability and that the scale may measure slightly different aspects of rater affect. Unfortunately, there has been very little research on how different components of affect might influence rating outcomes. Tsui and Barry (1986) used a measure of affect that consisted of three elements: admiration, respect, and liking. However, because they used an overall summary measure it was not possible to draw any conclusions about the relative contribution of each element. Further complexity is introduced by Murphy and Cleveland (1995) who have argued that undirected affective reactions, such as mood and temperament, may also influence rating outcomes. If it is acknowledged that affective reactions can

be complex and multifaceted, then future research on rater affect should be directed toward identifying critical components of affect, and determining how they interact to influence rating outcomes.

Rating Purpose

Contrary to the stated hypothesis, rating purpose did not influence rating outcomes in the present study. There was a slight tendency for raters completing evaluations for reference reports to be more lenient in their ratings than those completing evaluations for performance appraisals, but this difference was not significant. The failure to observe a significant difference for rating purpose may be the consequence of a weak experimental manipulation. There was a sizeable set of experimental materials that each participant was required to read and understand. Furthermore, the background information and instructions provided to participants in the two rating purpose conditions were nearly identical. Instructions to participants in the performance appraisal condition simply specified that ratings were being completed as part of a performance appraisal and that the university considered them an important part of staff development. Instructions to participants in the referee condition specified that ratings were for a reference report and that the university considered these an important part of the selection process. Under these conditions it is possible that the statement of rating purpose may not have been salient enough.

The influence of rating purpose on rating outcomes is thought to be a result of rater motivation (Murphy & Cleveland, 1995; Salvemini et al., 1993). Raters are assumed to be motivated to provide accurate ratings in some circumstances, but not others. In the case of reference reports, it seems that poor evaluations are likely to have negative consequences for the ratee. Raters, therefore, might be motivated to provide lenient ratings, and so avoid negative consequences for the ratee.

For performance appraisals the situation is more ambiguous. Performance appraisal information can be used for multiple and often conflicting purposes (Cleveland et al., 1989; Rudman, 1995). When appraisal information is to be used for administrative purposes, such as promotions or salary reviews, then the negative consequences of poor evaluations are readily apparent. However, appraisal information can also be used for developmental purposes (to improve employee performance), or simply for documentation reasons. Poor evaluations in appraisals collected for these latter purposes, and particularly for employee development, do not necessarily result in negative consequences for the appraisee. In fact, they can result in positive outcomes if employee performance improves as a consequence. Evaluations from raters in the performance appraisal condition in the present study might have varied depending upon what they perceived the eventual disposition of those ratings to be, and the consequences for the ratee. If they believed that ratings were to be used for administrative purposes, they may have been inclined to be more lenient than if they believed they were to be used for developmental purposes. This means that the strength of the experimental manipulation of purpose

may have been attenuated simply because instructions to the participants did not provide clear enough indications about the eventual disposition of their ratings.

Unfortunately, any explanations for the absence of an effect for rating purpose must remain conjectural at this stage. This is because the current study did not include a manipulation check. Hence, it is impossible to ascertain if raters were aware of the statement of rating purpose, or to what extent they may have been influenced by it. This is a major shortcoming and future studies should ensure that such checks are included as part of the experimental procedure.

Future research on rating purpose might wish to focus more closely on rating norms and on the perceived uses and consequences of evaluations. The central role of perceptions has been stressed by Ostroff (1993), who found that raters' beliefs about the purpose of appraisals influenced their rating behaviours, even when the explicit purpose of ratings did not vary. These results suggest that researchers might find it profitable to attend to the rater's interpretation or definition of the rating situation.

No studies to date have considered how referees' beliefs about the rating situation might influence rating outcomes. It is conceivable that leniency in reference reports may be at least partly attributable to normative beliefs among raters that discourage accuracy. Such norms might be based on the belief that the referee's primary obligation is to the applicant, and that ratings

are usually inflated. Certain features of reference reports may encourage and sustain such beliefs. Firstly, inflated ratings are in fact a common problem in reference reports (Muchinsky, 1979). Secondly, it is reasonable to assume that most applicants nominate referees that they believe will provide favourable evaluations. Moreover, many referees will have utilised a referee themselves at some stage during their career, and presumably would have selected someone who would rate them favourably. Therefore, referees will be aware of the expectation that they should be rating favourably and may feel obliged to do so.

The role of expectations could be tested by comparing ratings from referees nominated by applicants with those selected by the prospective employer. Any such study would have to control for length of acquaintance and knowledge of the applicant's performance, but would provide some insights about the strength of feelings of obligation, or reciprocity, that may contribute to the leniency effect.

Finally, as has been mentioned previously, referees may feel no special obligation to prospective employers to provide accurate ratings. Under such circumstances, norms fostering feelings of responsibility toward the applicant and endorsing leniency may well flourish and influence rating outcomes. Further investigations of rater accountability may help clarify this relationship, but in future studies, measures of rating norms should be included to assess the impact of normative beliefs and to help identify those situations in which they exert the greatest influence.

Ratee Characteristics

No significant differences in ratings as a function of ratee race or sex were found in the present study. The male lecturer described in the vignette was rated the same as the female lecturer, and likewise, the Asian lecturer was rated the same as the Caucasian lecturer. While the absence of any systematic rater bias based on ratee characteristics is encouraging, these findings are not consistent with recent studies which have reported that race and sex do influence ratings (Ford et al., 1986; Kraiger & Ford, 1985; Oppler et al., 1992; Pulakos et al., 1989). One explanation for this inconsistency is that the present study did not posses sufficient statistical power to detect effects that were present. Recent studies documenting the presence of rater bias have based their conclusions on the results from meta-analytic analyses (Ford et al., 1986; Kraiger & Ford, 1985), or from single studies with very large samples (Oppler et al., 1992; Pulakos et al., 1989). Furthermore, these studies have suggested that the magnitude of ratee race and sex effects is very small. Assuming a small effect size, a power analysis (Cohen, 1988) conducted on the present race and sex data (Total N = 288; p = .05; ES = .10) showed power was approximately 11%. As a consequence, the probability of a type 2 error, that is, claiming no effect when one was present, was close to 90%. The low level of statistical power in the current study makes it very difficult to interpret the null results associated with ratee race and sex.

The failure to detect any overt rater bias in the present study may be explicable on other grounds. One explanation is that bias is not dependent

solely on characteristics of the ratee, but instead, that it is a function of features of the rater, the ratee, and the rating situation. Studies have shown that these features can interact to produce bias. For example, Kraiger and Ford (1985) report a same-race bias in their meta-analysis of performance ratings. Barnes-Farrell, L'Heureux-Barrett, and Conway (1991) found that behaviours from female-typed task areas were rated more accurately in the context of female-typed occupations, and that behaviours from male-typed task areas were evaluated more accurately in the context of male-typed occupations. Rater bias may have been present, but remained undetected, because the simple analyses used in the current study assessed only the direct influence of ratee characteristics on rating outcomes.

In addition to examining how other variables might interact with ratee characteristics to influence rating outcomes, a worthwhile direction for future research would be to include a broader range of ratee characteristics. The present study was unusual in that it examined the possibility of a rater bias against Asians. In contrast, almost all of the published studies to date on race effects in performance ratings have compared Blacks and Whites. New Zealand is a multicultural society and the possibility of bias against members of other ethnic groups merits attention. Furthermore, race and sex are only two, out of a whole host of personal characteristics, that raters could use as cues to distort ratings when making performance judgements. Other factors such as disability, health status, age, and sexual orientation may also influence rating outcomes. Recent legislation in New Zealand (Employment Contracts Act, 1991; Human Rights Act, 1993) has identified a large number

of grounds on which it is illegal for employers to discriminate. This legislation would be a useful starting point for researchers who are concerned about the potential for characteristics of the ratee to affect rating outcomes.

Criterion Measures

Participants in the current study showed gains in accuracy for certain criterion measures, but not for others. This finding highlights two important points. The first of these is self-evident, that is, improvements in some components of accuracy are possible, while other components remain unaffected. The second point follows from the first, and is simply that conclusions about rating accuracy in research studies will be sensitive to the criterion measures selected. Therefore, it is important that researchers select measures appropriate to the phenomena under investigation.

The focus of the present study was on leniency in ratings. For investigators interested in leniency, McIntyre et al.'s (1984) leniency measure may be the most useful index of this particular outcome. This is because using other accuracy measures there is no way to tell if raters are being lenient or severe in their ratings. McIntyre et al.'s measure is much more descriptive and registers the tendency of raters to be favourable or harsh in their ratings. This is clear from the present study where the leniency measure produced a different pattern of results compared to the other accuracy measures that were calculated. However, the current results also suggest that a clearer

picture of rating accuracy can be obtained by employing multiple measures. Conclusions based on one measure can be misleading. In the present case there were differences between the three rating forms in their proneness to leniency errors. Although it is tempting to argue that asymmetrical scales produce the most accurate ratings, this does not appear to be the case. Measures of elevation, distance accuracy, and stereotype accuracy showed little differences between the forms. This finding suggests that reducing leniency in ratings does not necessarily improve raters' ability to distinguish between performance dimensions or result in gains in overall average accuracy.

Comparisons between accuracy and error measures produced results consistent, for the most part, with findings reported by Murphy and Balzer (1989). That is, error measures were found to be poor indicators of rating accuracy. In fact, two measures (Standard Deviation and Skewness) were found to be inversely related to rating accuracy. The presence of these rating errors was actually predictive of improvements in accuracy. With regards to these two particular measures, one must concur with recommendations from Murphy and Balzer that their use as indirect indicators of rating accuracy be discontinued. However, the remaining error measures (Mean and Lenmean) were found to be highly related to several of the accuracy measures. Murphy and Balzer did not calculate Lenmean values in their study, but did compute the error measure, Mean. Contrary to the results in the present study, they found that the value, Mean, was generally unrelated to the accuracy measures, and in fact, in some cases there was an inverse relationship. The

present results may be an artifact of the experimental design, in that the performance incidents depicted in the written vignette were selected to produce true performance levels as close to average as possible. Mean and Lenmean values, like the accuracy measures, are essentially deviation scores. However, rather than reflecting the distance of a rater's evaluation from a true score, they reflect the distance from the scale midpoint. Therefore, if scale midpoints and true scores are roughly equivalent, a high degree of relationship between the error measures and accuracy scores is to be expected.

Survey of Reference Reports

The current study shows that reference reports are viewed as an important part of the selection process. More than 90% of the participants indicated that they would request reference reports from applicants, and nearly half (49.5%) said they would not employ someone who had received a bad report. Experience as a manager, as a referee, and with selection in general, did not moderate respondents' perceptions regarding the value of reference reports. There was a clear consensus among participants that reference reports provide worthwhile information that assists employers engaged in selection decisions.

The present results are consistent with those reported in previous surveys. Researchers have repeatedly found that reference reports are very popular

and widely used by employers (Henderson, 1987; Mills, 1991; Patrickson & Haydon, 1988; Robertson & Makin, 1986; Vaughan & McLean, 1989). The value attached to reference reports by practitioners is surprising given their lacklustre record from studies that have evaluated their validity and reliability (see Chapter 2 for a review). If future surveys are to make any progress in resolving this apparent contradiction, they will have to consider a broader range of issues and do more than merely document the prevalence of reference reports. One way in which surveys can contribute is by identifying the characteristics of reference reports that make them so popular among employers. One possibility is that reference reports are viewed by employers as an inexpensive and relatively "hassle free" selection method. After all, it is the referees who shoulder most of the "work" involved in reference reports. In the final analysis, reference reports may be popular simply because they are convenient to use and entail very little effort on the part of the employer.

Limitations of the Current Study

There are several identifiable limitations of the present study that must be borne in mind when interpreting the results. One of the most critical of these relates to the nature of the stimuli on which ratings were based. Participants were required to evaluate a hypothetical lecturer whose performance was represented by statements contained in a written vignette. Using this method is an example of what Dipboye (1985) has called passive observer research.

The use of "paper people" as rating stimuli poses questions regarding the external validity of the research. Can the results from the present study be generalised to "real life" rating situations, such as performance appraisals or reference reports?

There is evidence that studies using "paper person" designs result in different experimental outcomes (larger effect sizes) compared to those that have used direct observation designs (Murphy, Herr, Lockhart, & Maguire, 1986). Woehr and Lance (1991) have tested competing explanations for these observed differences in effect sizes. They concluded that differences are attributable to a greater signal-to-noise ratio in direct observation studies. That is, effect sizes are smaller in direct observation studies because they include more performance irrelevant information (background noise) than do paper people studies. Interestingly, Woehr and Lance found that scripts in which performance statements were embedded in written descriptions that included irrelevant information resulted in recognition and accuracy rating outcomes similar to those obtained using videotape stimuli. They suggest that carefully constructed performance scripts can simulate some of the additional cues present in "real life" rating situations. However, they also point out that none of the laboratory methodologies, including those using direct observation techniques such as videotape, are likely to capture fully all aspects of the rating situation inherent in evaluations conducted in the "real world." Nevertheless, it has to be acknowledged that the present study represents an idealised rating situation where performance-irrelevant information has been minimised. Therefore, further research is required to establish if the results

can generalise to more complex and "noisier" environments found in applied rating situations.

A related issue concerns the nature of the rating task and the setting in which the study was conducted. Participants in the present investigation were students who were required to evaluate the performance of a lecturer. It has been suggested by some researchers that results from laboratory studies conducted in educational settings using upward appraisal may not generalise to different settings with non-student raters (Dipboye, 1985; Gordon, Slade, & Schmitt, 1986; Ilgen & Favero, 1985; Slade & Gordon, 1988). However, others have argued that the processes elucidated from laboratory research have external validity and that laboratory and field methodologies are complimentary (Dobbins, Lane, & Steiner, 1988a, 1988b; Mook, 1983; Woehr & Lance, 1991). Concerns that have been expressed regarding the generality of research findings are certainly reasonable. However, in the present case, characteristics of the sample may mitigate some of these concerns. More specifically, most of the distance education students who comprised the sample were experienced raters and were very familiar with reference reports and performance appraisals. Moreover, the majority of the sample were working full time, and, in addition, many of the participants were employed as managers or supervisors. The background and experience of the present sample sets them apart from the typical student participant used in many other investigations. In fact, their profile is likely to closely match that of raters in applied settings to whom the results are supposed to generalise. Nevertheless, limitations imposed by the artificial rating situation and nature

of the rating task remain, and place constraints on external validity in the present study.

In addition to the problem of external validity, there are several other methodological limitations that challenge the robustness of the results. For example, the manipulation of rating purpose was simplistic and poorly done. Part of the rationale for the manipulation was the avoidance of demand characteristics. However, in hindsight, a more thorough explanation of rating purpose would have been more likely to have communicated and established the desired motivational context. Another limitation was the fact that no manipulation checks were included. This omission means that it is difficult to determine if the failure to observe effects was due to the weak manipulation of the variable, failure to attend on the part of participants, or simply because the variable was irrelevant. One might also ask questions about the reliability of the measure of rater affect. Unfortunately, because it was a single item measure, no reliability coefficients could be calculated.

The low return rate in the present study is also of concern. Requests for participation were sent out to more than 900 individuals. Slightly less than 300 participated in the study, a return rate of only 31%. In hindsight it would have been worthwhile to include a follow-up letter which may have helped to bolster participant numbers. However, while the return rate was low it must be pointed out that it is consistent with those reported in other studies which have used postal surveys (e.g., Cleveland et al., 1989; Judge, Cable, Boudreau, & Bretz, 1995; Lyn, Cao, & Horn, 1996; Pazy, 1996).

Furthermore, many investigations have reported far lower return rates (e.g., Arthur & Bennett, 1995; Lin, 1996; Shaw, Kirkbride, Fisher, & Tang, 1995). Nevertheless, because of the low return rate the representativeness of the sample cannot be guaranteed, and questions remain concerning the external validity of the results. Finally, the investigation would have been improved if participants had been required to evaluate more than one ratee. The inclusion of multiple ratees would have resulted in a design that allowed for the calculation of the entire range of accuracy measures.

Future Research

For practitioners and researchers in the United States, any discussion of future research on reference reports may well be premature. The poor performance of reference reports that has been documented in the research literature, coupled with the current litigious climate prevailing in the United States, would appear to be a strong disincentive for referees to provide reference checks. Employers will be reluctant to evaluate previous employees if their judgements can be contested in a court of law with substantial payments in damages at stake. Ryan and Lasek (1991) have examined two areas of employer liability in relation to pre-employment inquiries in the United States, namely negligent hiring and defamation. They point out that these two legal doctrines pose a quandary for employers and may work at cross-purposes to subvert the selection process. On the one hand, the law relating to negligent hiring emphasises the importance of securing relevant

background information prior to hiring, particularly for jobs in which the employer may be held to have a "special duty of care". On the other hand, many employers are unwilling to provide such information for fear of breaching privacy regulations and/or incurring defamation suits. Recent privacy legislation introduced by the New Zealand government (Privacy Act, 1993) raises a similar spectre here. The special nature of reference checks (and other types of pre-employment inquiries) compounds this dilemma because, according to Ryan and Lasek, they function largely as screen-out procedures during selection. Screen-out procedures focus on negative selection, that is, identifying reasons to disqualify an individual from consideration for employment. Because screen-out procedures seek negative information (such as a lack of requisite knowledge, skills or abilities, or the presence of "character flaws" that may make misconduct more likely) they are clearly vulnerable to accusations of slander, libel, adverse impact and breaches of privacy. Ryan and Lasek note that the situation at present is one where employers are in the position of wanting information from others who are unwilling to provide that information.

Before an organisation introduces any selection procedure a thorough job analysis is recommended, if not essential. In the United States, job analysis represents the standard used to judge the content validity of the assessment procedure and provides the basis from which criteria are developed for the purposes of assessing criterion-related validity (U.S. Equal Employment Opportunity Commission et al., 1978, 1979). In New Zealand there are no legal requirements for organisations to conduct job analyses, but they do

constitute "best practice." If reference reports are based on a job analysis, and there is evidence of improved validity, then there is probably no reason why there should be any special legal impediments preventing employers from using reference checks. Similar sentiments are expressed by Ryan and Lasek (1991) in their review of pre-employment inquiries. They stress the importance of comprehensive job analysis procedures and the linkage of knowledge, skills and abilities to selection methods. Employers who adopt such procedures should be able to reduce the likelihood of negligent hiring suits. Job analysis is also central to the development of sound, behaviourally based performance appraisal systems. Such systems can provide concrete evidence to substantiate negative statements made by referees and, consequently, may be of benefit in avoiding potential claims of defamation.

Alternative methods for collecting information from referees warrants additional study. The telephone check in particular has been under-researched, yet would appear to have considerable promise. Behaviourally oriented methods, such as situational interviewing (Latham et al., 1980) and the accomplishment record (Hough, 1984) offer appropriate models on which to base the telephone check. Using job analysis information, employers can ask applicants to nominate referees who are able to supply examples of the applicants' achievements in critical areas. The referees can then be contacted and the information elicited. The referee is no longer responsible for providing ratings, thus minimizing leniency errors, and at the same time guaranteeing the job-relatedness of the information. Furthermore, organisations would

then be able to capitalise on the advantages of appropriate training and selection.

The incorporation of job analyses and the utilization of behaviourally based items, or scales, would be expected to enhance reliability and validity of reference reports. Whether or not this is the case is an empirical question to be addressed by future research. Another avenue researchers may wish to explore is methods for developing accountability on the part of referees. Mero and Motowildo (1995) found that participants who were made to feel accountable by having to justify their ratings to the experimenter provided more accurate ratings of a simulated subordinate. Requiring referees to justify their evaluations might result in similar improvements. Simply providing feedback on the accuracy of their judgements, or even the knowledge that such feedback will be provided, could also result in improvements in the quality of the information supplied by referees. Clear specifications by employers as to who are acceptable referees, and the kind of information they will be asked to supply, may also help ensure greater consistency across applicants.

The significance of information sources has been noted by Murphy and Cleveland (1995). They identified a number of potential sources who could supply performance appraisal information. Murphy and Cleveland categorised these sources according to their hierarchical position within the organisation, and their likely access to information about behaviours and the results of behaviours related to task accomplishment and interpersonal

relations at work. Table 21 is adapted from Murphy and Cleveland. It uses the same categorisation scheme but, in this case, it is applied to likely sources of information for reference reports. The material in Table 21 suggests that employers must be careful to target appropriate referees if they wish to obtain information germane to task accomplishment and interpersonal relations in the work environment. For example, relatives and friends have very little access to information about an individual's work-related task and interpersonal behaviours and results. On the other hand, subordinates, peers, and the immediate supervisor all have much greater access to such information. The question of which source, or combination of sources, that can provide the most valid ratings merits further attention.

Table 21

Referees' access to information about task and interpersonal behaviours and results (Adapted from Murphy & Cleveland, 1995)

Source	Task		Interpersonal	
	Behaviours	Results	Behaviours	Results
Relatives	Rare	Rare	Rare	Rare
Friends	Rare	Rare	Rare	Rare
Subordinates	Rare	Occasional	Frequent	Frequent
Peers	Frequent	Frequent	Frequent	Frequent
Immediate Supervisor	Occasional	Frequent	Occasional	Occasional
Upper Management	Rare	Occasional	Rare	Rare

Researchers should also consider how extraneous factors related to the referee influence the employer's interpretation of evaluations contained in reference reports. For example, does the age, sex, race, or status of the

referee have any bearing on the decisions reached by employers? In a similar vein, investigation of the criteria used by applicants when they select referees would also be worthwhile. What are the characteristics of referees that applicants value, and why do they choose some people to be referees, and not others?

The nature of the information contributed by reference reports is also worthy of consideration. Is this information unique, or do other selection procedures with superior predictive validity (such as structured interviews, cognitive ability tests, biodata) render the reference report redundant? Multiple assessments will be required to answer this, and to establish if reference reports do have any incremental validity. Although there are many directions in which future research on reference reports can head, studies that explicitly address their deficiencies must be a priority. Only when the deficiencies are remedied can their full potential be realised.

Summary and Conclusions

Although the present research program suffered from some methodological limitations, several conclusions can be drawn. Firstly, the meta-analysis showed that changes in rating format, and specific efforts to reduce leniency in ratings, are associated with notable increases in the validity coefficients obtained from reference reports. More specifically, it showed that the predictive validity of structured reference reports is superior to that of

unstructured reports. However, substantial variance remains unaccounted for and this indicates that other factors may also moderate the predictive validity of reference reports. Nonetheless, organisations that collect ratings from referees as part of their selection process are advised to make use of structured forms based on job analysis information if they wish to maximise the predictive validity of their reference reports.

Secondly, the present study is the first to show that the introduction of asymmetrical, positively toned rating scales reduces leniency in ratings when evaluating others. These scales were found to reduce leniency in evaluations for reference reports and performance appraisals compared to ratings collected using Likert-type scales, or those based on narrative comments from raters. While Likert-type scales produced ratings with less leniency than evaluations based on narrative comments, they realised smaller gains than the asymmetrical, positively toned rating forms. Asymmetrical, positively toned scales offer a cheap and easily implemented method for controlling leniency in ratings gathered from referees. The whole-hearted endorsement of this rating method should be tempered by the recognition that reductions in leniency may not produce similar improvements in other types of rating accuracy.

Thirdly, rater affect was found to be associated with leniency in evaluations. Raters who expressed greater liking for the ratee tended to be more favourable in their assessments, whereas raters who disliked the ratee tended to be more severe in their ratings. These results emphasise the difficulties confronting

employers who seek objective evaluations of applicants and current employees. Research on techniques that can assist raters to separate subjective feelings from objective judgements is called for. Structured rating forms derived from job analysis information may be helpful in this respect by focussing raters' attention on job-related performance dimensions. Further investigations broadening the conceptual and operational definitions of rater affect would also be worthwhile.

Fourthly, there was no evidence that the ratees' sex or race influenced rating outcomes. The absence of any systematic rater bias of this type is encouraging. However, researchers are advised to be vigilant and to continue with investigations that examine the impact of extraneous ratee characteristics on evaluations by raters. Bias in ratings is, to some extent, a product of contemporary social conditions. As conditions change, and society redefines itself, new biases may arise, or old familiar ones may become re-established.

Finally, evidence from the survey included in the second part of the research program shows that reference reports are very popular, and are thought to contribute vital information to the selection process. This raises concerns that practitioners are unaware of the empirical evidence that indicates that reference reports are poor predictors of job performance. More alarming is the possibility that practitioners are aware of this evidence, yet choose to ignore it. Although a great deal more research on methods for improving the validity of reference reports is required, care must be taken to disseminate the

findings in such a way that they are accessible to practitioners, and encourage compliance with "best practice." It is also essential that future surveys establish how employers use reference reports. Further studies directed toward improving the predictive validity of information from referees may prove redundant if reference reports are used primarily as a screen out procedure.

The premise on which reference reports are based, the use of a third party to supply a sample of behaviour regarding an applicant, seems intrinsically sound, but only if the sample helps to predict subsequent behaviour on the job. Perhaps the best way of insuring that this occurs is to conduct a thorough job analysis. This will provide the basis for relevant questions which can then be directed toward knowledgeable referees. In such circumstances, reference reports may yet prove to be one of the most cost-effective approaches to personnel selection available.

References

Arthur, W., Jr., & Bennett, W., Jr. (1995). The international assignee: The relative importance of factors perceived to contribute to success. *Personnel Psychology, 48*, 99-114.

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86*, 736-765.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-322.

Arvey, R. D., Miller, H. E., Gould, R., & Burch, P. (1987). Interview validity for selecting sales clerks. *Personnel Psychology, 40*, 1-12.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72*, 567-572.

Balzer, W. (1986). Biases in the recording of performance-related information: The effects of initial impression and centrality of the appraisal task. *Organizational Behavior and Human Decision Processes, 37*, 329-347.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38*, 335-345.

Bannister, B. D., Kinicki, A. J., DeNisi, A. S., & Hom, P. W. (1987). A new method for the statistical control of rating error in performance ratings. *Educational and Psychological Measurement, 47*, 583-596.

Barnes-Farrell, J. L., L'Heureux-Barrett, T. J., & Conway, J. M. (1991). Impact of gender-related job features on the accurate evaluation of performance information. *Organizational Behavior and Human Decision Processes, 48*, 23-35.

Bartol, K. M., & Butterfield, D. A. (1976). Sex effects in evaluating leaders. *Journal of Applied Psychology, 61*, 446-454.

Bass, B. M. (1956). Reducing leniency in merit ratings. *Personnel Psychology, 9*, 359-369.

Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology, 66*, 296-301.

Beason, G. M., & Belt, J. A. (1976). Verifying applicant's backgrounds. *Personnel Journal, 55*, 345-348.

Berkshire, J. R., & Highland, R. W. (1953). Forced-choice performance rating-A methodological study. *Personnel Psychology, 6*, 356-378.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*, 301-308.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training. New response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60-66.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard rating scale: A new rating system. *Personnel Psychology, 25*, 185-199.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80*, 168-177.

Borresen, H. A. (1967). The effects of instructions and item content on three types of ratings. *Educational and Psychological Measurement, 27*, 855-862.

Browning, R. C. (1968). Validity of reference ratings from previous employers. *Personnel Psychology, 21*, 389-393.

Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70*, 108-115.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*, 25-42.

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.

Carroll, S. J., & Nash, A. N. (1972). Effectiveness of a forced-choice reference check. *Personnel Administration, 35*, 42-46.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among thorns? *Personnel Psychology, 32*, 751-766.

Ceci, S. J., & Peters, D. (1984). Letters of reference: A naturalistic study of the effects of confidentiality. *American Psychologist, 39*, 29-31.

Centra, J. A. (1976). The influence of different directions on student ratings of instruction. *Journal of Educational Measurement, 13*, 277-282.

Champion, C. H., Green, S. B., & Sauser, W. I. (1988). Development and evaluation of shortcut-derived behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*, 29-41.

Chen, M. (1993). Discrimination in New Zealand: A personal journey. In E. McDonald and G. Austin (Eds.), *Claiming the law. Essays by New Zealand women in celebration of the 1993 suffrage centennial.* Wellington: Victoria University Press.

Cleveland, J. N., & Landy, F. J. (1981). The influence of rater and ratee age on two performance judgements. *Personnel Psychology, 34*, 19-29.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130-135.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaun.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115*, 465-474.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218-244.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177-193.

Decotiis, T. A., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review, 3*, 635-646.

Dickinson, T. L., & Glebocki, G. G. (1990). Modifications in the format of the mixed standard scale. *Organizational Behavior and Human Decision Processes, 47*, 124-137.

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

Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resumé evaluations. *Journal of Applied Psychology, 62*, 288-294.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1986). The effects of ratee sex and purpose of appraisals on the accuracy of performance evaluations. *Basic and Applied Social Psychology, 7*, 225-241.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology, 73*, 551-558.

Dobbins, G. H., Lane, I. M., & Steiner, D. D. (1988a). A note on the role of laboratory methodologies in applied behavioural research: Don't throw the baby out with the bath water. *Journal of Organizational Behavior, 9*, 281-286.

Dobbins, G. H., Lane, I. M., & Steiner, D. D. (1988b). A further examination of student babies and laboratory bath water: A response to Slade and Gordon. *Journal of Organizational Behavior, 9*, 377-378.

Doverspike, D., Cellar, D. F., & Hajek, M. (1987). Relative sensitivity to performance cue effects as a criterion for comparing rating scale formats. *Educational and Psychological Measurement, 47*, 1135-1139.

Downie, N. M., & Heath, R. W. (1965). *Basic statistical methods.* (2nd ed.). New York: Harper & Row and John Weatherhill Inc.

Driscoll, L. A., & Goodwin, W. L. (1979). The effects of varying information about use and disposition of results on university students' evaluations of faculty and courses. *American Educational Research Journal, 16*, 25-37.

Duarte, N. T., Goodson, J. R., & Klich, N. R. (1993). How do I like thee? Let me appraise the ways. *Journal of Organizational Behavior, 14*, 239-249.

Duarte, N. T., Goodson, J. R., & Klich, N. R. (1994). Effects of dyadic quality and duration on performance appraisal. *Academy of Management Journal, 37*, 499-521.

Employment Contracts Act, 1991, No.22. Wellington, New Zealand: Government Printer.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology, 35*, 105-116.

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

Ferguson, L. W. (1949). The value of acquaintance ratings in criterion research. *Personnel Psychology, 2*, 93-102.

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.

Festinger, L. (1954). *A theory of cognitive dissonance*. Evanston, IL: Row Peterson.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203-210.

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

Fox, S., Caspy, T., & Reisler, A. (1994). Variables affecting leniency, halo and validity of self-appraisal. *Journal of Occupational and Organizational Psychology, 67*, 45-56.

Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology, 53*, 518-524.

Goheen, H. W., & Mosel, J. N. (1959). Validity of the employment recommendation questionnaire: II. Comparison with field investigation. *Personnel Psychology, 12*, 297-301.

Gomez-Mejia, L. R. (1988). Evaluating employee performance: Does the appraisal instrument make a difference? *Journal of Organizational Behavior Management, 9*, 155-172.

Gordon, M. E., Slade, L. E., & Schmitt, N. (1986). The "science of sophomore" revisited: From conjecture to empiricism. *Academy of Management Review, 11*, 191-207.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59*, 705-711.

Härtel, C. E. J. (1993). Rating format research revisited: Format effectiveness and acceptability depend on rater characteristics. *Journal of Applied Psychology, 78*, 212-217.

Hauenstein, N. M. A. (1992). An information-processing approach to leniency in performance judgements. *Journal of Applied Psychology, 77*, 485-493.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*, 68-73.

Henderson, R. H. (1987). *Investigation into staff selection practices in Canterbury businesses.* Unpublished manuscript, Massey University, Department of Management and Administration, New Zealand.

Highhouse, S. (1992). The leniency scale: Is it really independent of ratee behavior? *Educational and Psychological Measurement, 52*, 781-786.

Hough, L. M. (1984). Development and evaluation of the "Accomplishment Record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69*, 135-146.

Human Rights Act 1993, No. 82. Wellington, New Zealand: Government Printer.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-99.

Hunter, J. E., & Schmidt, F. L. (1977). A critical analysis of the statistical and ethical implications of various definitions of test fairness. *Psychological Bulletin, 83*, 1053-1071.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Beverly Hills, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54*, 321-368.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10*, 311-321.

Izraeli, D. N., & Izraeli, D. (1985). Sex effects in evaluating leaders. *Journal of Applied Psychology, 70*, 540-546.

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgement decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology, 75*, 500-505.

Jones, A., & Harrison, E. (1982). Prediction of performance in initial officer training using reference reports. *Journal of Occupational Psychology, 55*, 35-42.

Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz, R. D., Jr. (1995). An empirical investigation of the predictors of executive career success. *Personnel Psychology, 48*, 483-519.

Judge, T. A., & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal, 36*, 80-105.

Kane, J. S. (1983). Performance distribution assessment: A new breed of appraisal methodology. In H. J. Bernardin & R. W. Beatty (Eds.), *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* (3rd ed.). Englewood Cliffs: Prentice-Hall.

Kingston, N. (1971). *Selecting managers: A survey of current practice in 200 companies.* London: British Institute of Management.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology, 34*, 263-289.

Kingstrom, P. O., & Mainstone, L. E. (1985). An investigation of rater-ratee acquaintance and rater bias. *Academy of Management Journal, 28*, 641-653.

Kinicki, A. J., & Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*, 17-27.

Kinicki, A. J., Hom, P. W., Trost, M. R., & Wade, K. J. (1995). Effects of category prototypes on performance-rating accuracy. *Journal of Applied Psychology, 80*, 354-370.

Knight, F. B. (1923). The effect of the "acquaintance factor" upon personnel judgements. *Journal of Educational Psychology, 14*, 129-142.

Knouse, S. B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology, 36*, 331-341.

Kornhauser, A. W. (1927). A comparison of raters. *Personnel Journal, 5*, 338-344.

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

Kryger, B. R., & Shikiar, R. (1978). Sexual discrimination in the use of letters of recommendation: A case of reverse discrimination. *Journal of Applied Psychology, 63*, 309-314.

Krzystofiak, F., Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology, 73*, 515-521.

Landy, F. J. (1989). *Psychology of work behavior* (4th ed.). Pacific Grove, CA: Brooks/Cole.

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

Latham, G. P. (1986). Job performance and appraisal. In C. L. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology*. London: Wiley.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*, 255-268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*, 550-555.

Lin, C. Y. Y. (1996). Training and development practices in Taiwan: A comparison of Taiwanese, American and Japanese firms. *Asia Pacific Journal of Human Resources, 34*, 26-43.

Longenecker, C. O., & Gioia, D. A. (1992). The executive appraisal paradox. *Academy of Management Executive, 6*, 18-28.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1*, 183-193.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*, 66-71.

Lynn, S. A., Cao, L. T., & Horn, B. C. (1996). The influence of career stage on the work attitudes of male and female accounting professionals. *Journal of Organizational Behavior, 17*, 135-139.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment, 1*, 18-26.

Maurer, T. J., & Alexander, R. A. (1991). Contrast effects in behavioral measurement: An investigation of alternative process explanations. *Journal of Applied Psychology, 76*, 3-10.

Maurer, T. J., & Taylor, M. A. (1994). Is sex by itself enough? An exploration of gender bias issues in performance appraisal. *Organizational Behavior and Human Decision Processes, 60*, 231-251.

Mayfield, E. C. (1964). The selection interview: A reevaluation of published research. *Personnel Psychology, 17*, 239-260.

McGregor, J., Thomson, M., & Dewe, P. (1994). Women in management in New Zealand: A benchmark survey. *Women in Management: Series Paper No 19*, Faculty of Commerce, University of Western Sydney, Nepean.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147-156.

Meier, R. S., & Feldhusen, J. F. (1979). Another look at Dr. Fox: Effect of stated purpose for evaluation, lecturer expressiveness, and density of lecture content on student ratings. *Journal of Educational Psychology, 71*, 339-345.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*, 517-524.

Mills, A. J. (1991). *Personnel consulting firms' managerial selection methods.* Unpublished Master's thesis, University of Waikato, Hamilton, New Zealand.

Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field study of adverse impact and generalizability. *Academy of Management Journal, 25*, 598-606.

Mohrman, A. M., & Lawler, E. E. (1983). Motivation and performance appraisal behavior. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory.* Hillsdale, NJ: Lawrence Erlbaum.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.

Mosel, J. N., & Goheen, H. W. (1952). Agreement among replies to an employment recommendation questionnaire. *American Psychologist, 7*, 365-366.

Mosel, J. N., & Goheen, H. W. (1958a). The validity of the employment recommendation questionnaire in personnel selection: 1. Skilled traders. *Personnel Psychology, 11*, 481-490.

Mosel, J. N., & Goheen, H. W. (1958b). Use of the ERQ in hiring. *Personnel Journal, 36*, 338-340.

Mosel, J. N., & Goheen, H. W. (1959). The employment recommendation questionnaire: III. Validity of different types of references. *Personnel Psychology, 12*, 469-477.

Muchinsky, P. M. (1979). The use of reference reports in personnel selection: A review and evaluation. *Journal of Occupational Psychology, 52*, 287-297.

Murphy, K. R. (1991). Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy. *Organizational Behavior and Human Decision Processes, 50*, 45-50.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619-624.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of purpose of rating on accuracy in observing teacher behavior and evaluating teacher performance. *Journal of Educational Psychology, 76*, 45-54.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal Of Applied Psychology, 70*, 72-84.

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

Murphy, K. R., & Constans, J. I. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology, 72*, 573-577.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*, 320-325.

Murphy, K. R., Herr, B. M., Lockhart, M. C., & Maguire, E. (1986). Evaluating the performance of paper people. *Journal of Applied Psychology, 71*, 654-661.

Murphy, K. R., Philbin, T. A., & Adams, S. R. (1989). Effect of purpose of observation on accuracy of immediate and delayed performance ratings. *Organizational Behavior and Human Decision Processes, 43*, 336-354.

Napier, N., & Latham, G. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology, 39*, 827-837.

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. *Journal of Applied Psychology, 68*, 102-114.

Newman, S. H., & Howell, M. A. (1961). Validity of forced-choice items for obtaining references on physicians. *Psychological Reports, 8*, 367.

Nieva, V. F., & Gutek, B. A. (1980). Sex effects on evaluation. *Academy of Management Review, 5*, 267-276.

Norton, S. D., Gustafson, D. P., & Foster, C. E. (1977). Assessment for management potential: Scale design and development, training effects and rater/ratee sex effects. *Academy of Management Journal, 20*, 117-131.

Norusis, M. J. (1992). *SPSS/PC+ Base system users guide version 5.0.* Chicago: SPSS Inc.

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*, 201-217.

Ostroff, C. (1993). Rater perceptions, satisfaction and performance ratings. *Journal of Occupational and Organizational Psychology, 66*, 345-356.

Padgett, M. Y., & Ilgen, D. R. (1989). The impact of ratee performance characteristics on rater cognitive processes and alternative measures of rater accuracy. *Organizational Behavior and Human Decision Processes, 44*, 232-260.

Parsons, C. K., & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology, 69*, 557-568.

Patrickson, M., & Haydon, D. (1988). Management selection practices in South Australia. *Human Resource Management Australia, 26*, 96-104.

Paunonen, S. V., Jackson, N. D., & Oberman, S. M. (1987). Personnel selection decisions: Effects of applicant personality and the letter of reference. *Organizational Behaviour and Human Decision Processes, 40*, 96-114.

Pazy, A. (1986). The persistence of pro-male bias despite identical information regarding causes of success. *Organizational Behavior and Human Decision Processes, 38*, 366-377.

Pazy, A. (1996). Concept and career-stage differentiation in obsolescence research. *Journal of Organizational Behavior, 17*, 59-78.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373-406.

Peres, S. H., & Garcia, R. (1962). Validity and dimensions of descriptive adjectives used in reference letters for engineering applicants. *Personnel Psychology, 15*, 279-286.

Petzel, T. P., & Berndt, D. J. (1980). APA internship selection criteria: Relative importance of academic and clinical preparation. *Professional Psychology, 11*, 792-796.

Privacy Act 1993, No. 28. Wellington, New Zealand: Government Printer.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38*, 76-91.

Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of Management Journal, 26*, 129-139.

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

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76*, 432-446.

Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on *p* variables. *Journal of Educational Statistics, 7*, 139-156.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 33*, 1-62.

Rhea, B. D. (1966). *Validation of OCS selection instruments: The relationship of OCS selection measures to OCS performance.* U.S. Naval Personnel Research Activity, Technical Bulletin STB 66-18, San Diego, CA.

Rhea, B. D., Rimland, B., & Githens, W. H. (1965). *The development and evaluation of a forced-choice letter of reference form for selecting officer candidates.* U.S. Naval Personnel Research Activity, Technical Bulletin STB 66-10, San Diego, CA.

Robbins, T. L., & DeNisi, A. S. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology, 79*, 341-353.

Robertson, I. T., & Makin, P. J. (1986). Management selection in Britain: A survey and critique. *Journal of Occupational Psychology, 59*, 45-57.

Rudman, R. (1995). *Performance planning and review: Making employee appraisals work.* Melbourne: Pitman.

Ryan, A. M., & Lasek, M. (1991). Negligent hiring and defamation: Areas of liability related to pre-employment inquiries. *Personnel Psychology, 44*, 293-319.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 55*, 41-60.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697-798.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial setting. *Journal of Applied Psychology, 57*, 237-241.

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

Schneider, D. J. (1991). Social cognition. *Annual Review of Psychology, 42*, 527-561.

Schriesheim, C. A., Kinicki, A. J., & Schriesheim, J. F. (1979). The effect of leniency on leader behavior descriptions. *Organizational Behavior and Human Performance, 23*, 1-29.

Schwab, D. P., Heneman, H. G., & Decotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

Shaffer, D. R., Mays, P. V., & Etheridge, K. (1976). Who shall be hired: A biasing effect of the Buckley amendment on employment practices? *Journal of Applied Psychology, 61*, 571-575.

Shaffer, D. R., & Tomarelli, M. (1981). Bias in the ivory tower: An unintended consequence of the Buckley amendment for graduate admissions? *Journal of Applied Psychology, 66*, 7-11.

Sharon, A. T. (1970). Eliminating bias from student ratings of college instructors. *Journal of Applied Psychology, 54*, 278-281.

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 22*, 251-263.

Shaw, J. B., Kirkbride, P. S., Fisher, C. D., & Tang, S. F. Y. (1995). Human resource practices in Hong Kong and Singapore: The impact of political forces and imitation processes. *Asia Pacific Journal of Human Resources, 33*, 22-39.

Shilton, J. D., McGregor, J., & Tremaine, M. (in press). Feminising the boardroom. *Women in Management Review, 11*.

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

Slade, L. A., & Gordon, M. E. (1988). On the virtues of laboratory babies and student bath water: A reply to Dobbins, Lane, and Steiner. *Journal of Organizational Behavior, 9*, 373-376.

Sleight, R. B., & Bell, G. D. (1954). Desirable content of letters of recommendation. *Personnel Journal, 32*, 421-422.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*, 22-40.

Smith, M., & George, D. (1992). Selection methods. *International Review of Industrial and Organisational Psychology, 7*, 55-97.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Smither, J. W., Reilly, R. R., & Buda, R. (1988). Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology, 73*, 487-496.

Spitzform, M., & Hamilton, S. (1976). A survey of directors from APA approved internship programs on intern selection. *Professional Psychology, 7*, 406-410.

Spriegel, W. R., & James, V. A. (1958). Trends in recruitment and selection practices. *Personnel, 35*, 42-48.

Stedman, J. M., Costello, R. M., Gaines, T. Jr., Schoenfeld, L. S., Loucks, S., & Burstein, A. G. (1981). How clinical psychology interns are selected: A study of decision-making processes. *Professional Psychology, 12*, 415-419.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*, 136-142.

Steiner, D. D., Rain, J. S., & Smalley, M. M. (1993). Distributional ratings of performance: Further examination of a new rating format. *Journal of Applied Psychology, 78*, 438-442.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506.

Sulsky, L. M., & Day, D. V. (1992). Frame of reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*, 501-510.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper Row.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4*, 39-47.

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*, 747-753.

Tsui, A. S., & Barry, B. (1986). Interpersonal affect and rating errors. *Academy of Management Journal, 29*, 586-599.

Tsui, A. S., & O'Reilly III, C. A. (1989). Beyond simple demographic effects: The importance of relational demography in superior-subordinate dyads. *Academy of Management Journal, 32*, 402-423.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, J. G. Ghurye, W. Hoeffding, W. G. Madoo, & H. Mann (Eds.), *Contributions to probabilities and statistics: Essays in honor of Harold Hoetelling.* Stanford, CA: Stanford University Press.

Turban, D. B., & Jones, A. P. (1988). Supervisor-subordinate similarity: Types, effects, and mechanisms. *Journal of Applied Psychology, 73*, 228-234.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.

U. S. Equal Employment Opportunity Commission, U. S. Civil Service Commission, U. S. Department of Labor, & U. S. Department of Justice (1978). *Uniform Guidelines on Employment Selection Procedures.* Federal Register, 43, 166 38290-38309.

U. S. Equal Employment Opportunity Commission, U. S. Civil Service Commission, U. S. Department of Labor, & U. S. Department of Justice (1979). *Adoption of questions and answers to clarify and provide a common interpretation of the Uniform Guidelines on Employment Selection Procedures.* Federal Register, 44, 167 11996-12009.

Vaughan, E., & McLean, J. (1989). A survey and critique of management selection practices in Australian business firms. *Asia Pacific Human Resource Management, 27,* 20-33.

Wagner, R. (1949). The employment interview: A critical review. *Personnel Psychology, 2,* 17-46.

Waldman, D. A., & Avolio, B. (1986). A meta-analysis of age differences in job performance. *Journal of Applied Psychology, 71,* 33-38.

Wayne, S. J., & Ferris, G. R. (1990). Influence tactics, affect, and exchange quality in supervisor-subordinate interactions: A laboratory experiment and field study. *Journal of Applied Psychology, 75,* 487-499.

Wayne, S. J., & Kacmar, K. M. (1991). The effects of impression management on the performance appraisal process. *Organizational Behavior and Human Decision Processes, 48,* 70-88.

Wayne, S. J., & Liden, R. C. (1995). Effects of impression management on performance ratings: A longitudinal study. *Academy of Management Journal, 38,* 232-260.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72*, 484-487.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples and criteria. *Journal of Applied Psychology, 52*, 372-376.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*, 315-321.

Wiesner, W. A., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Decision Processes, 35*, 314-339.

Woehr, D. J. (1994). Understanding frame-of-reference-training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79*, 525-534.

Woehr, D. J., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgement in performance appraisal: The tip of the iceberg. *Journal of Applied Psychology, 78*, 232-241.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Woehr, D. J., & Lance, C. E. (1991). Paper people versus direct observation: An empirical examination of laboratory methodologies. *Journal of Organizational Behavior, 12*, 387-397.

Wright, O. R., Jr. (1969). Summary of research on the selection interview since 1964. *Personnel Psychology, 22*, 391-413.

Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62*, 191-199.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151-175.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*, 117-124.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67*, 752-758.

Appendix 1

Letter Requesting Participation in the Study

Dear Student,

My name is Karl Pajo and I am a staff member in the Department of Human Resource Management at Massey University. I am asking you to participate in a study I am undertaking as part of my PhD research. The focus of the study is on how we make judgements about human performance, particularly in the context of referee reports.

WHAT WILL YOU HAVE TO DO?

If you are willing to participate please fill in the consent form at the end of this letter. You should then read the vignette describing the job performance of a university lecturer that accompanies this posting. You should rate how likable you consider the person described in the vignette to be, using the rating scale provided at the end of the vignette. You should tear off the consent form and then place all of the materials in the envelope provided so that they are ready for posting back to Massey University. It is important that the consent form, likability rating and vignette all be returned to me.

You are then ready to proceed to the next part of the study. Included in the materials that have been posted to you are two brief questionnaires. The first asks you some information about yourself (e.g., have you ever provided a referee report) while the second asks for some ratings or comments about the person described in the vignette. **It is very important that you do not refer back to the vignette when writing your comments or ratings.** The study is not concerned about recall of information but rather the impressions you form about people.

Once you have finished place all of the materials in the envelope and return them to me at Massey University.

HOW MUCH TIME WILL THIS TAKE?

The entire exercise should take no more than 15-20 minutes. This includes the time taken to read the vignette and complete all the ratings. It may take a little longer if there is not a postbox conveniently nearby.

WHY SHOULD I DO THIS?

There are a number of worthwhile reasons for being involved in this research. Firstly, you will be contributing in a very concrete way to our understanding of the performance appraisal process. Performance appraisal is a critical part of many managers' jobs and any advances in this area would have important practical implications for employers, employees and organisations. Furthermore, as a small country we often rely on overseas findings which may not be appropriate for our

circumstances. This research project is particularly valuable because it is being conducted here in New Zealand.

Participation in research is a valuable learning and educational experience and is an integral part of your total university schooling. Moreover, it may help to make many of the articles you read more concrete and understandable, and at the very least, should give you more of a feel for the research process.

Finally, this research project depends on a large sample of subjects. I would be very grateful and you would be doing me a considerable favour by agreeing to participate.

WHAT HAPPENS TO THE DATA?

All information is handled only by me. No subject will be identifiable in the results or in the raw data once questionnaires have been returned. A summary of the results will be made available to any subject who wishes a copy. Your willingness to participate in this study in no way affects your grades in any university paper.

If you have any questions at all regarding this research please feel free to contact me, or my supervisor, here at the university.

Karl Pajo
Department of Human Resource Management
Massey University
Private Bag
Palmerston North
ph (06) 350 4283
email *K.B.Pajo@massey.ac.nz*

Dr. John Podd
Department of Psychology
Massey University
Private Bag
Palmerston North
ph (06) 350 4135
email *J.V.Podd@massey.ac.nz*

Once again I encourage you to participate in the research and offer my thanks to all who choose to do so.

Yours sincerely

Karl Pajo

CONSENT FORM

I _____ (print your name) agree to participate in Karl Pajo's PhD research on performance judgements .

Date: _____ Signature: _____

Appendix 2

Teaching Vignette and Likability Scale

VIGNETTE

Mary Goh is a lecturer at the university. She is Chinese and is thirty years old. She has been employed by the university for the last four years. Each year the students' association and the university administration ask students to comment on the performance of the lecturers they come into contact with. The following excerpts are statements from these annual surveys describing the job performance of Mary Goh. You should take these excerpts as being representative of her everyday performance on the job.

Please read the excerpts carefully. Once you have finished you will be asked some questions about them.

The lecturer never praises or offers encouragement to the class.

The lecturer used numerous visual aids, handouts, and examples to illustrate her lectures.

The lecturer's eyes light up when she discusses the material.

The lecturer distributed the workload evenly throughout the course.

The lecturer made a fool out of a student in class for asking a ridiculous question.

The lecturer would sometimes come to lectures and forget her overheads.

The lecturer relied heavily on her notes, thus making very little eye contact with her students.

The lecturer clearly set out all the due dates for assignments, required readings, textbooks, office hours, location of tutorials, lecture times in the course materials provided at the start of the year.

The lecturer always kept her classroom presentations specific and to the point.

The lecturer tested us on material she did not cover.

The lecturer travels in order to see and hear things about her profession which she then shares with her students.

The lecturer never made an effort to speak to anyone in class.

The lecturer requires a lot of memorization for her class.

The lecturer always acted excited and happy to be in class.

The lecturer always told you well in advance when assignments were due.

The lecturer does not scale grades unless the class does really badly.

The lecturer gives partial credit if she can see that you are on the right track.

The lecturer often told the class about interesting articles she had read or experiments she had heard about.

The lecturer gives hard tests which require the students to study a lot.

The lecturer gives students her office number but does not make them feel welcome.

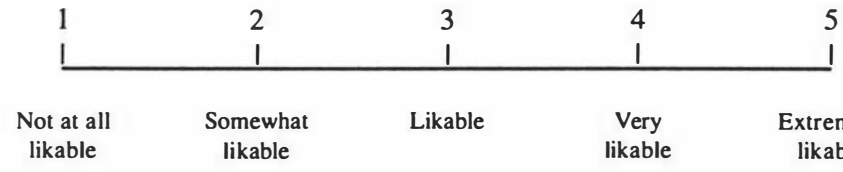
The lecturer mistakenly scheduled due dates for some assignments on public holidays.

The lecturer takes into consideration students' other classes and outside activities when assigning work.

The lecturer mumbled during her lectures.

The lecturer gives more notes in one hour than most lecturers do in two.

Based on the descriptions of the lecturer you have read so far, and using the scale presented below, please rate how likable you find her. Write your rating in the box provided on the right.



You should now place the signed consent form and this vignette into the envelope provided so that they are ready to be returned to Massey University. Once you have done that you can proceed to the next part of the study, that is, completing the demographic questionnaire and the rating form. **Remember, you should not refer back to this vignette when completing the rating form.**

VIGNETTE

Mary Keppel is a lecturer at the university. She is Pakeha and is thirty years old. She has been employed by the university for the last four years. Each year the students' association and the university administration ask students to comment on the performance of the lecturers they come into contact with. The following excerpts are statements from these annual surveys describing the job performance of Mary Keppel. You should take these excerpts as being representative of her everyday performance on the job.

Please read the excerpts carefully. Once you have finished you will be asked some questions about them.

The lecturer never praises or offers encouragement to the class.

The lecturer used numerous visual aids, handouts, and examples to illustrate her lectures.

The lecturer's eyes light up when she discusses the material.

The lecturer distributed the workload evenly throughout the course.

The lecturer made a fool out of a student in class for asking a ridiculous question.

The lecturer would sometimes come to lectures and forget her overheads.

The lecturer relied heavily on her notes, thus making very little eye contact with her students.

The lecturer clearly set out all the due dates for assignments, required readings, textbooks, office hours, location of tutorials, lecture times in the course materials provided at the start of the year.

The lecturer always kept her classroom presentations specific and to the point.

The lecturer tested us on material she did not cover.

The lecturer travels in order to see and hear things about her profession which she then shares with her students.

The lecturer never made an effort to speak to anyone in class.

The lecturer requires a lot of memorization for her class.

The lecturer always acted excited and happy to be in class.

The lecturer always told you well in advance when assignments were due.

The lecturer does not scale grades unless the class does really badly.

The lecturer gives partial credit if she can see that you are on the right track.

The lecturer often told the class about interesting articles she had read or experiments she had heard about.

The lecturer gives hard tests which require the students to study a lot.

The lecturer gives students her office number but does not make them feel welcome.

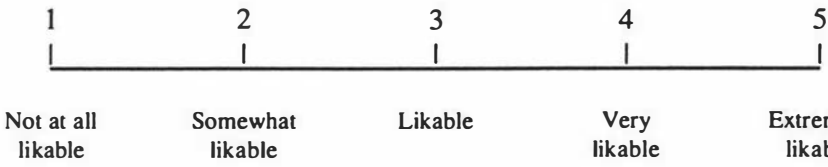
The lecturer mistakenly scheduled due dates for some assignments on public holidays.

The lecturer takes into consideration students' other classes and outside activities when assigning work.

The lecturer mumbled during her lectures.

The lecturer gives more notes in one hour than most lecturers do in two.

Based on the descriptions of the lecturer you have read so far, and using the scale presented below, please rate how likable you find her. Write your rating in the box provided on the right.



You should now place the signed consent form and this vignette into the envelope provided so that they are ready to be returned to Massey University. Once you have done that you can proceed to the next part of the study, that is, completing the demographic questionnaire and the rating form. **Remember, you should not refer back to this vignette when completing the rating form.**

VIGNETTE

John Goh is a lecturer at the university. He is Chinese and is thirty years old. He has been employed by the university for the last four years. Each year the students' association and the university administration ask students to comment on the performance of the lecturers they come into contact with. The following excerpts are statements from these annual surveys describing the job performance of John Goh. You should take these excerpts as being representative of his everyday performance on the job.

Please read the excerpts carefully. Once you have finished you will be asked some questions about them.

The lecturer never praises or offers encouragement to the class.

The lecturer used numerous visual aids, handouts, and examples to illustrate his lectures.

The lecturer's eyes light up when he discusses the material.

The lecturer distributed the workload evenly throughout the course.

The lecturer made a fool out of a student in class for asking a ridiculous question.

The lecturer would sometimes come to lectures and forget his overheads.

The lecturer relied heavily on his notes, thus making very little eye contact with his students.

The lecturer clearly set out all the due dates for assignments, required readings, textbooks, office hours, location of tutorials, lecture times in the course materials provided at the start of the year.

The lecturer always kept his classroom presentations specific and to the point.

The lecturer tested us on material he did not cover.

The lecturer travels in order to see and hear things about his profession which he then shares with his students.

The lecturer never made an effort to speak to anyone in class.

The lecturer requires a lot of memorization for his class.

The lecturer always acted excited and happy to be in class.

The lecturer always told you well in advance when assignments were due.

The lecturer does not scale grades unless the class does really badly.

The lecturer gives partial credit if he can see that you are on the right track.

The lecturer often told the class about interesting articles he had read or experiments he had heard about.

The lecturer gives hard tests which require the students to study a lot.

The lecturer gives students his office number but does not make them feel welcome.

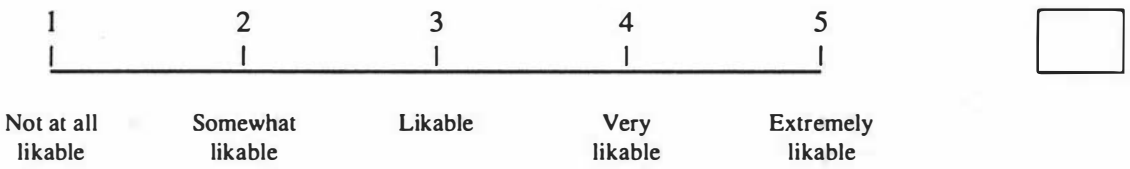
The lecturer mistakenly scheduled due dates for some assignments on public holidays.

The lecturer takes into consideration students' other classes and outside activities when assigning work.

The lecturer mumbled during his lectures.

The lecturer gives more notes in one hour than most lecturers do in two.

Based on the descriptions of the lecturer you have read so far, and using the scale presented below, please rate how likable you find him. Write your rating in the box provided on the right.



You should now place the signed consent form and this vignette into the envelope provided so that they are ready to be returned to Massey University. Once you have done that you can proceed to the next part of the study, that is, completing the demographic questionnaire and the rating form. **Remember, you should not refer back to this vignette when completing the rating form.**

VIGNETTE

John Keppel is a lecturer at the university. He is Pakeha and is thirty years old. He has been employed by the university for the last four years. Each year the students' association and the university administration ask students to comment on the performance of the lecturers they come into contact with. The following excerpts are statements from these annual surveys describing the job performance of John Keppel. You should take these excerpts as being representative of his everyday performance on the job.

Please read the excerpts carefully. Once you have finished you will be asked some questions about them.

The lecturer never praises or offers encouragement to the class.

The lecturer used numerous visual aids, handouts, and examples to illustrate his lectures.

The lecturer's eyes light up when he discusses the material.

The lecturer distributed the workload evenly throughout the course.

The lecturer made a fool out of a student in class for asking a ridiculous question.

The lecturer would sometimes come to lectures and forget his overheads.

The lecturer relied heavily on his notes, thus making very little eye contact with his students.

The lecturer clearly set out all the due dates for assignments, required readings, textbooks, office hours, location of tutorials, lecture times in the course materials provided at the start of the year.

The lecturer always kept his classroom presentations specific and to the point.

The lecturer tested us on material he did not cover.

The lecturer travels in order to see and hear things about his profession which he then shares with his students.

The lecturer never made an effort to speak to anyone in class.

The lecturer requires a lot of memorization for his class.

The lecturer always acted excited and happy to be in class.

The lecturer always told you well in advance when assignments were due.

The lecturer does not scale grades unless the class does really badly.

The lecturer gives partial credit if he can see that you are on the right track.

The lecturer often told the class about interesting articles he had read or experiments he had heard about.

The lecturer gives hard tests which require the students to study a lot.

The lecturer gives students his office number but does not make them feel welcome.

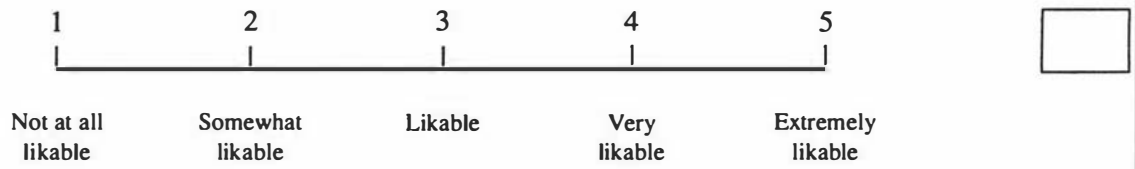
The lecturer mistakenly scheduled due dates for some assignments on public holidays.

The lecturer takes into consideration students' other classes and outside activities when assigning work.

The lecturer mumbled during his lectures.

The lecturer gives more notes in one hour than most lecturers do in two.

Based on the descriptions of the lecturer you have read so far, and using the scale presented below, please rate how likable you find him. Write your rating in the box provided on the right.



You should now place the signed consent form and this vignette into the envelope provided so that they are ready to be returned to Massey University. Once you have done that you can proceed to the next part of the study, that is, completing the demographic questionnaire and the rating form. **Remember, you should not refer back to this vignette when completing the rating form.**

Appendix 3

Demographic Questionnaire

DEMOGRAPHIC QUESTIONNAIRE

Name: _____ Date of Birth: _____

Current Job Title: _____

Ethnicity: _____ Sex: Male/Female (please circle)
(e.g. Chinese, Maori, Pakeha etc)

Q1. Have you ever attended any university lectures? Yes / No (please circle)

Q2. How many university papers have you completed? No. = ☐

Q3. Have you ever carried out a performance appraisal? Yes / No (please circle)

If YES, then approximately how many times? No. = ☐

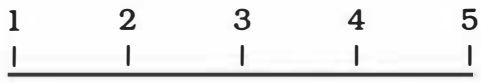
Q4. Have you ever been responsible for selecting new staff? Yes / No (please circle)

If YES, then approximately how many employment decisions have you been involved in? No. = ☐

Q5. Would you ask for referee reports when selecting staff? Yes / No (please circle)

Q6. Would you ever hire anyone who had a bad referee report? Yes / No (please circle)

Q7. Using the following scale how useful do you believe referee reports are? Rating = ☐



Not at all Useful | Somewhat Useful | Useful | Very Useful | Extremely Useful

Q8. Have you ever acted as a referee for someone? Yes / No (please circle)

If YES, then approximately how many times? No. = ☐

Appendix 4

Unstructured Rating Forms Developed for the Referee's Report and for the Performance Appraisal

RATING FORM I

John Goh has applied for a teaching position in the university and has given your name as a referee. The university considers referee reports to be a very important part of the selection process and would value your opinion on this candidate's suitability for the appointment.

Please comment on the applicant's teaching ability. When framing your remarks you should comment on each of the following elements;

1. The applicant's ability to organise and plan.
2. The applicant's interest in the material they teach.
3. The applicant's relations with students.
4. The applicant's ability to deliver lectures.
5. The applicant's ability to assess work fairly.
6. The applicant's ability to provide a reasonable workload for students.
7. The applicant's overall teaching ability.

Organise & Plan:

Interest in Material:

Relations with Students:

Deliver Lectures:

Assess Work Fairly:

Workload:

Overall Teaching Ability:

RATING FORM I

John Goh's annual performance appraisal is due and you have been asked to provide some information. The university considers performance appraisals to be a very important part of staff development and would value your opinion on this staff member's abilities.

Please comment on the lecturer's teaching ability. When framing your remarks you should comment on each of the following elements;

1. The lecturer's ability to organise and plan.
2. The lecturer's interest in the material they teach.
3. The lecturer's relations with students.
4. The lecturer's ability to deliver lectures.
5. The lecturer's ability to assess work fairly.
6. The lecturer's ability to provide a reasonable workload for students.
7. The lecturer's overall teaching ability.

Organise & Plan:

Interest in Material:

Relations with Students:

Deliver Lectures:

Assess Work Fairly:

Workload:

Overall Teaching Ability:

Appendix 5

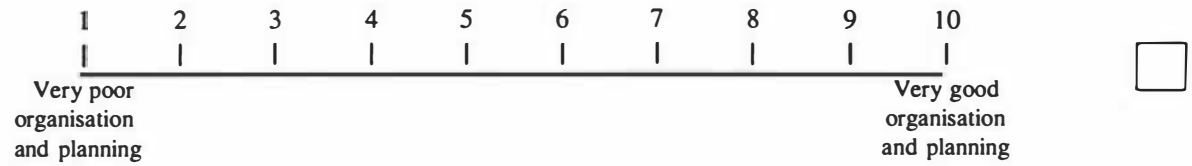
Likert-Type Rating Forms Developed for the Referee's Report and for the Performance Appraisal

RATING FORM II

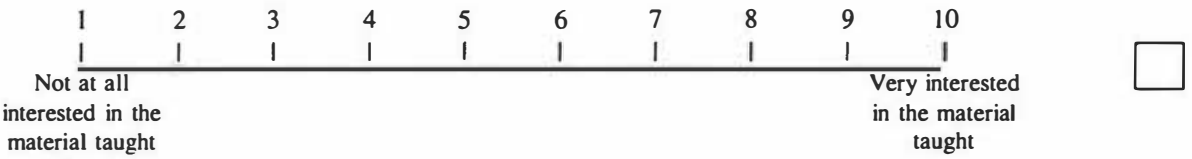
Mary Keppel has applied for a teaching position in the university and has given your name as a referee. The university considers referee reports to be a very important part of the selection process and would value your opinion on this candidate's suitability for the appointment.

Please rate the applicant's teaching ability on the following scales. You should place your rating in the box provided to the right of each scale. Use only whole numbers for your rating.

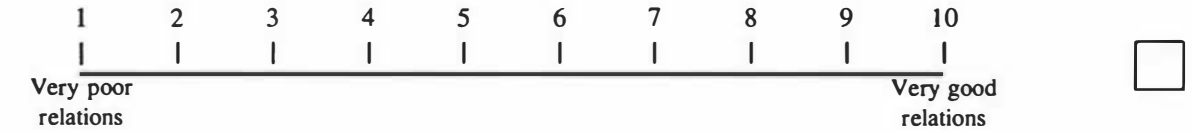
1. The applicant's ability to organise and plan.



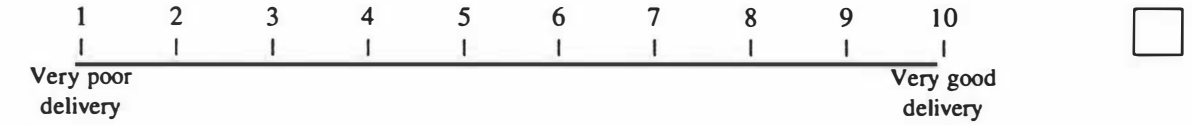
2. The applicant's interest in the material taught.



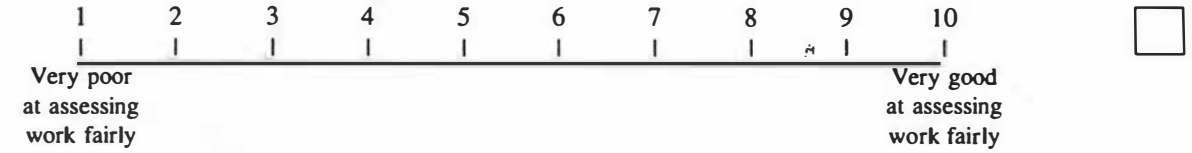
3. The applicant's relations with students.



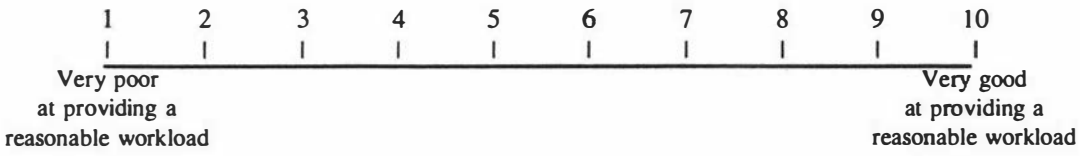
4. The applicant's ability to deliver lectures.



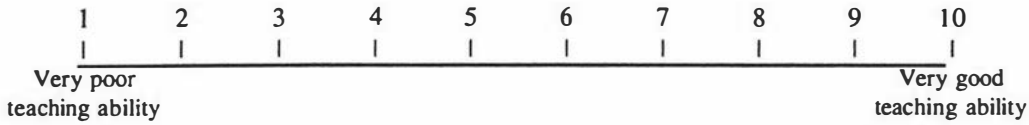
5. The applicant's ability to assess work fairly.



6. The applicant's ability to provide a reasonable workload for students.



7. The applicant's overall teaching ability.



RATING FORM II

Mary Keppel's annual performance appraisal is due and you have been asked to provide some information. The university considers performance appraisals to be a very important part of staff development and would value your opinion on this staff member's abilities.

Please rate the lecturer's teaching ability on the following scales. You should place your rating in the box provided to the right of each scale. Use only whole numbers for your rating.

1. The lecturer's ability to organise and plan.

1 2 3 4 5 6 7 8 9 10 ☐

Very poor organisation and planning — Very good organisation and planning

2. The lecturer's interest in the material taught.

1 2 3 4 5 6 7 8 9 10 ☐

Not at all interested in the material taught — Very interested in the material taught

3. The lecturer's relations with students.

1 2 3 4 5 6 7 8 9 10 ☐

Very poor relations — Very good relations

4. The lecturer's ability to deliver lectures.

1 2 3 4 5 6 7 8 9 10 ☐

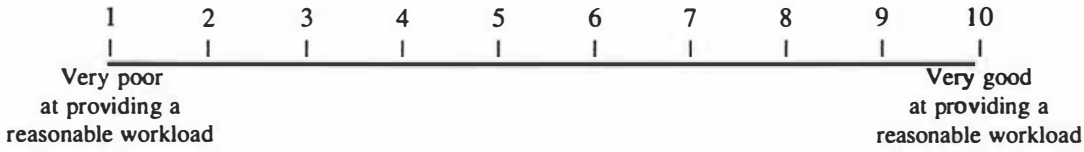
Very poor delivery — Very good delivery

5. The lecturer's ability to assess work fairly.

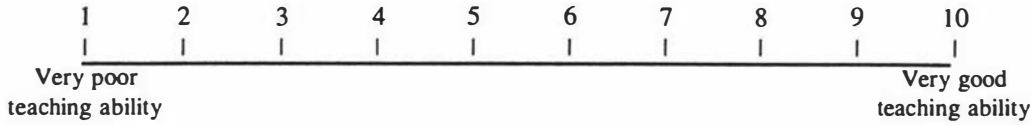
1 2 3 4 5 6 7 8 9 10 ☐

Very poor at assessing work fairly — Very good at assessing work fairly

6. The lecturer's ability to provide a reasonable workload for students.



7. The lecturer's overall teaching ability.



Appendix 6

Asymmetrical Rating Forms Developed for the Referee's Report and for the Performance Appraisal

RATING SCALE III

John Keppel has applied for a teaching position in the university and has given your name as a referee. The university considers referee reports to be a very important part of the selection process and would value your opinion on this candidate's suitability for the appointment.

Please rate the applicant's teaching ability on the following scales. You should rate the applicant by placing a tick in the appropriate box.

1. The applicant's ability to organise and plan.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

2. The applicant's interest in the material taught.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

3. The applicant's relations with students.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

4. The applicant's ability to deliver lectures.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

5. The applicant's ability to assess work fairly.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

6. The applicant's ability to provide a reasonable workload for students.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

7. The applicant's overall teaching ability.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

RATING SCALE III

John Keppel's annual performance appraisal is due and you have been asked to provide some information. The university considers performance appraisals to be a very important part of staff development and would value your opinion on this staff member's abilities.

Please rate the lecturer's teaching ability on the following scales. You should rate the lecturer by placing a tick in the appropriate box.

1. The lecturer's ability to organise and plan.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

2. The lecturer's interest in the material taught.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

3. The lecturer's relations with students.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

4. The lecturer's ability to deliver lectures.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

5. The lecturer's ability to assess work fairly.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

6. The lecturer's ability to provide a reasonable workload for students.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

7. The lecturer's overall teaching ability.

☐ Poor ☐ Ok ☐ Satisfactory ☐ Good ☐ Very Good ☐ Excellent

Appendix 7

Gender-Typed Versions of the Leniency Scale

Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether it is likely to be true or false as it pertains to the individual described in the vignette. You will not have firsthand knowledge of these attitudes and traits, so simply use your overall impression of the person described in the vignette to decide whether it is true or false. Check each item in the true or false column, but not both. Do not omit or skip any items. **You should complete your ratings without referring back to the vignette.**

	TRUE	FALSE
1. She's always willing to admit it when she makes a mistake.	____	____
2. She always tries to practice what she preaches.	____	____
3. She doesn't seem to find it difficult to get along with loud mouthed, obnoxious people.	____	____
4. She sometimes tries to get even, rather than forgive and forget.	____	____
5. When she doesn't know something, she doesn't mind at all admitting it.	____	____
6. She is always courteous, even to people who are disagreeable.	____	____
7. At times she has really insisted on having things her own way.	____	____
8. She would never think of letting someone else be blamed for her mistakes.	____	____
9. She never hesitates to go out of her way to help people in trouble.	____	____
10. She has never shown intense dislike for anyone.	____	____
11. She sometimes seems resentful when she doesn't get her way.	____	____
12. She is always careful about her manner of dress.	____	____
13. Her social manners are always perfect.	____	____
14. If she could get something without paying for it and be sure that she was not seen, she would probably do it.	____	____
15. She likes to gossip at times.	____	____
16. No matter who she's talking to, she's always a good listener.	____	____
17. There have been occasions when she took advantage of someone.	____	____
18. She never resents being asked to return a favour.	____	____
19. She never gets irked when people express ideas very different from her.	____	____
20. There are times when she seems to get quite jealous of the good fortune of others.	____	____
21. She almost never tells someone off.	____	____
22. She sometimes gets irritated by people who ask favours of her.	____	____
23. She has never deliberately said something that hurt someone's feelings.	____	____

Appendix 7

Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether it is likely to be true or false as it pertains to the individual described in the vignette. You will not have firsthand knowledge of these attitudes and traits, so simply use your overall impression of the person described in the vignette to decide whether it is true or false. Check each item in the true or false column, but not both. Do not omit or skip any items. **You should complete your ratings without referring back to the vignette.**

	TRUE	FALSE
1. He's always willing to admit it when he makes a mistake.	_____	_____
2. He always tries to practice what he preaches.	_____	_____
3. He doesn't seem to find it difficult to get along with loud mouthed, obnoxious people.	_____	_____
4. He sometimes tries to get even, rather than forgive and forget.	_____	_____
5. When he doesn't know something, he doesn't mind at all admitting it.	_____	_____
6. He is always courteous, even to people who are disagreeable.	_____	_____
7. At times he has really insisted on having things his own way.	_____	_____
8. He would never think of letting someone else be blamed for his mistakes.	_____	_____
9. He never hesitates to go out of his way to help people in trouble.	_____	_____
10. He has never shown intense dislike for anyone.	_____	_____
11. He sometimes seems resentful when he doesn't get his way.	_____	_____
12. He is always careful about his manner of dress.	_____	_____
13. His social manners are always perfect.	_____	_____
14. If he could get something without paying for it and be sure that he was not seen, he would probably do it.	_____	_____
15. He likes to gossip at times.	_____	_____
16. No matter who he's talking to, he's always a good listener.	_____	_____
17. There have been occasions when he took advantage of someone.	_____	_____
18. He never resents being asked to return a favour.	_____	_____
19. He never gets irked when people express ideas very different from his.	_____	_____
20. There are times when he seems to get quite jealous of the good fortune of others.	_____	_____
21. He almost never tells someone off.	_____	_____
22. He sometimes gets irritated by people who ask favours of him.	_____	_____
23. He has never deliberately said something that hurt someone's feelings.	_____	_____

Appendix 8

Comparison of Rating Form Accuracy Using the Original Rating Metric

Table 22
Results of t-tests comparing mean accuracy values for ratings from three different forms using the original rating metric

Accuracy Measure	Mean Scores		*t* value	*df*	Significance
	Form 1	*Form 2*			
Elevation	1.16	0.87	2.75	178.00	< .008
Stereotype Accuracy	1.60	1.34	4.16	178.00	< .001
Distance Accuracy	1.72	1.42	4.19	178.00	< .001
Leniency	0.83	0.42	2.62	178.00	<.008
	Form 1	*Form 3*			
Elevation	1.16	0.67	5.26	115.62	< .001
Stereotype Accuracy	1.60	1.09	10.12	128.17	< .001
Distance Accuracy	1.72	1.07	10.61	133.54	< .001
Leniency	0.83	-0.04	5.94	132.39	< .001
	Form 2	*Form 3*			
Elevation	0.87	0.67	2.56	178.06	*ns*
Stereotype Accuracy	1.34	1.09	4.98	170.60	< .001
Distance Accuracy	1.42	1.07	6.00	183.50	< .001
Leniency	0.42	-0.04	3.60	196.23	<.001

a - Lower values denote greater accuracy.
Critical alpha p < .008.