ASSESSING DISCRIMINATION IN A POLICE RECRUIT ASSESSMENT CENTER

by

Paul N. Tinsley

B.A., Simon Fraser University, 1977
M.A., Simon Fraser University, 1988

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF EDUCATION
IN EDUCATIONAL LEADERSHIP AND POLICY

in

THE FACULTY OF GRADUATE STUDIES

(Department of Educational Studies)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Studies

The University of British Columbia
Vancouver, Canada

Date 2000 Sep 26

ABSTRACT

The overall concern of this study is that of substantive equality, as defined by Canadian law, in the employment context, and the specific goal of this study is to provide a model to assess (and prevent) unlawful systemic discrimination in an assessment center. Because discrimination is essentially the same, wherever it occurs, the model proposed in this study is also useful for assessing discrimination in employment selection generally.

In the employment context, evidence of systemic discrimination is often limited to selection patterns, and so this study argues that statistical analyses can be particularly useful. Since the Supreme Court adopted the effects theory, where intent is immaterial and the focus is on results, such analyses are likely to become an appealing alternative to traditional arguments of exclusion and disproportion. The analytic model proposed here suggests two general phases to a legal analysis of discrimination. First, there is the preliminary phase, which consists of three interrelated steps: identifying the applicable selection procedure, identifying the relevant legal issue, and identifying the appropriate groups for comparison. Second, there is the assessment phase, which consists of two sequential steps: comparing the groups of interest on the dimension of interest to determine if differences exist, and analyzing observed differences to determine if they are legally or practically significant. It is in this phase that statistical analyses can be especially helpful in an assessment of systemic discrimination.

To test its utility, the proposed model was applied to the Justice Institute of British Columbia Police Academy assessment center (where entry level police applicants are screened) to determine whether the assessment center discriminated on the basis of sex. Of particular interest to the Police Academy is that the results indicated no sex

discrimination, but notably the results also indicated that the proposed model provides a practicable and relatively uncomplicated way to assess discrimination. Moreover, consistent with the goal of prevention, this study demonstrates how a reliability assessment can provide important information about the potential for discrimination in employee selection, thereby providing employers with the means to be more proactive than otherwise possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Carolyn Shields, the Chair of my Research Supervisory Committee, and Dr. Tom Sork, both of the Department of Educational Studies, The University of British Columbia, and Dr. Darryl Plecas, of the Department of Criminal Justice, University College of the Fraser Valley, for their guidance and for permitting me to pursue this study. In addition, I wish to express my appreciation to Mr. Steve Watt, the Director of the Justice Institute of British Columbia (JIBC) Police Academy, Mr. Bob Hull, the former Director (retired), and Sgt. Marilyn MacDonald, the Administrator of the JIBC Police Academy assessment center, for their assistance and support. And finally, I wish to express my appreciation to Mr. Mark Brosinski, a University College of the Fraser Valley student, who patiently entered 22 years of data into SPSS. Without the help of each one of these people, and especially the encouragement of Dr. Darryl Plecas, I would never have been able to complete this study. Thank you.

DEDICATION

To my wife, Nancy:

Though all the sky be parchment

And ink be found in oceans blue

This all would be insufficient

To write of my love for you

CHAPTER I: INTRODUCTION

Employers are always interested in new assessment[1] procedures that promise to identify the most qualified candidates for a target job or position, and the police are no exception. There is no question that an organization, especially one such as policing that requires complex skills, cannot survive without personnel selection procedures that have the ability to differentiate between and among candidates. But while this ability is essential for efficient and effective personnel selection, in the absence of proactive measures to guard against differentiating on prohibited grounds such as sex or race, unlawful discrimination is almost inevitable. Twenty years ago Cronbach, Yalow and Schaeffer (1980) cautioned that "the fairness of procedures used to choose applicants should be examined closely for legal and practical reasons" (p. 693), but their caution has been largely ignored.

With this in mind, the overall concern of this study is that of substantive equality, as defined by Canadian law, in the employment context. Specifically, the goal of this study is to provide a model or framework in which unlawful systemic discrimination can be assessed and, as a result, prevented within an assessment center.[2] To test the utility of the proposed model, it will be operationally applied to the Justice Institute of British Columbia (JIBC)[3] Police Academy assessment center to determine whether the

---

[1] In employment selection literature, the term "assessment" has a broad definition that includes terms such as "test," "scale," "instrument," or any other technique by which measurements are obtained on candidates in order to differentiate between them on some dimension of interest.

[2] The term "assessment center" refers to the particular standardized assessment methodology that has been approved by the International Congress on the Assessment Center Method (Guidelines and Ethical Considerations, 1989; see also Standards and Ethical Considerations, 1975 & 1979). The assessment center method will be discussed in detail in Chapter 2.

[3] The JIBC is an umbrella organization for seven relatively autonomous academies related to justice and public safety, of which the Police Academy is one. Hereafter, the JIBC Police Academy will be referred to only as the Police Academy.

assessment center discriminates on the prohibited ground of sex.[4] In so doing, the proposed model will be evaluated, the Police Academy will benefit from the results of the study, and a contribution to the literature on personnel selection will be made.

Being an introduction, this chapter only provides the bare bones version of the proposed model, but the analysis of the Police Academy provides an opportunity to flesh out the model as the study progresses. That said, this chapter also provides background and context, discusses why the problem of discrimination is a pressing issue, and explains the organization of the study.

Background and Context

In 1977 the British Columbia Police Commission,[5] after reviewing the selection and promotional procedures used by municipal police, supported an initiative to implement a relatively new and innovative procedure known as the assessment center method in an attempt "to improve the quality of policing" in the province (Turner & Higgins, 1977, p. 1; see also Turner, 1978; and Sgt. W. Bryant, Vancouver Police Department, personal communication, February 2, 1977).[6] What distinguishes the assessment center method from traditional pencil and paper tests is that it relies on

[4] Because there is an insufficient number of female raters to permit an analysis of interaction effects between rater sex and candidate sex, the analysis is limited to "main effects" (i.e., whether assessment center scores are a function of candidate sex). And because no practical method is available to determine the race of either raters or candidates, the analysis is limited to sex discrimination. Women, however, have been traditionally excluded from policing and so an analysis of this variable is important and may indicate whether further research is warranted.

[5] The British Columbia Police Commission was legislated out of existence on July 1, 1998 (Police Act, 1996, updated to include amendments enacted 1997-37-1 to 46 effective July 1, 1998; rf. Police Amendment Act, 1997).

[6] In 1994 The Policing in British Columbia Commission of Inquiry (commonly known as the "Oppal Report") also recommended the assessment center method, stating that it was a personnel selection system that could help police identify candidates who would likely make a positive contribution to the community (p. E-33).

observed behavior rather than written responses. Basically, an assessment center consists of a team of trained raters who systematically rate a candidate's performance in job-related simulations to determine his or her suitability for the target job or position.

In 1978 the Police Academy, in partnership with the municipal policing community and the provincial Police Commission, contracted with Development Dimensions International (DDI)[7] to establish an assessment center. Originally, the main purpose of this assessment center was to identify candidates who exhibited the greatest potential for supervisory and management positions; however, over time the emphasis shifted to identifying recruit candidates who exhibited the greatest potential for entry level positions into policing (British Columbia Police Commission, 1996, p. 68; McClellan, 1985, p. 1; Taylor, 1983, p. 45).

Discrimination in Policing

This section begins with a discussion about organizational complacency and the problem of identifying discrimination (both of which exacerbate the general problem of discrimination), and ends with a discussion of sex discrimination in policing.

Organizational Complacency

Although the Police Commission was the government agency responsible for policy relating to police administration, no policy was ever drafted to guide the operations of the new assessment center.[8] In 1990, a Police Commission working

[7] DDI, founded by D.W. Bray and W.C. Byham, is an international company that consults in the field of human resources. Bray and Byham are also generally recognized as the founders of the contemporary assessment center (discussed in Chapter 2), and Byham played an important role in establishing the Police Academy assessment center.

[8] As a result of a Police Commission initiative, prescribed standards for training at the Police Academy were established by regulation in the Rules Regarding Training in 1981, but no standards were prescribed for the assessment center.

committee recommended minimum provincial standards for municipal police (British Columbia Police Commission, 1990 & 1995), which included a recommendation that the Police Academy follow the "Guidelines and Ethical Considerations" (1989) of the International Congress on the Assessment Center Method.[9] The Guidelines represent an attempt by a loosely knit group of human resource professionals (including Bray and Byham (see footnote 7)) to establish minimum operating standards for assessment centers. Among other things, these standards require written policy on almost all aspects of an assessment center operation, including training, rater selection, and the assessment of discrimination and validity.

Despite the recommendation by the Police Commission and the standards endorsed by the Guidelines, and despite being the "gatekeeper" for municipal police, to this day the Police Academy has never written policy for the administration of the assessment center, nor has it ever proactively addressed the issue of discrimination in the assessment process. If, as Singer (1993) argues, an organization's structure and procedures are manifestations of its values (p. 2), then it seems that the Police Academy has fallen short of its ethical and legal obligations to ensure that its practice is consistent with its principles.[10]

---

[9] These minimum standards were subsequently endorsed in 1992 by "order of the lieutenant governor in council" (Provincial Standards for Municipal Police Departments in British Columbia). In 1994 The Policing in British Columbia Commission of Inquiry (i.e., "Oppal Report") recommended to the Attorney General that the standards applying to the assessment center be made mandatory (p. E-35), but this was never done.

[10] According to the philosophy statement found in the Provincial standards for municipal police departments in British Columbia, the police must seek and preserve public favour by exercising impartial service, "without regard to race, national or ethnic origin, colour, religion, sex, belief, or social standing" (British Columbia Police Commission, 1995).

Identifying Discrimination

Compounding the problem of organizational complacency, which is not limited to the Police Academy,[11] is the problem of identifying discrimination. In an organization, apart from obvious cases of direct discrimination (e.g., where a company standard or policy overtly excludes females),[12] cases of indirect or systemic discrimination are not readily apparent because they are usually concealed by ostensibly neutral policies and practices.[13] To make matters worse, discrimination has received relatively little attention in personnel selection literature in general (as noted by Schmitt, Gooding, Noe & Kirsch, 1984, pp. 407, 420; Singer, 1993, pp. 1-3, 16, 29) and in assessment center literature in particular (as noted by Hoffman & Thornton, 1997, p. 456; Lowry, 1993, p. 489; and Singer, 1993, p. 31), which tend to focus on issues of validity and utility. As argued by Feltham (1988), who was commenting on police assessment centers, a great deal of time, effort and resources are spent on traditional concerns while a basic human rights issue such as discrimination is ignored (p. 142).

In Canada the situation is no better, where discrimination has received relatively little attention in personnel selection literature in general (as noted by Singer, 1993, p. 30) and in police selection literature in particular (as noted by Nelson, 1992, p. 193). In policing, because of its high profile in the area of social order, there has been a strong call

[11] For example, see "A Pale, Male Reflection of the Community: Equity Still Eludes Toronto Police" (1999) and "Proposed Guidelines for Recruitment and Selection of Visible Minority Police Officers in Canada" (1986); see also Nelson (1992).

[12] See O'Malley v. Simpson-Sears (1985), where the Supreme Court provided a specific example (p. 551). Types of discrimination will be discussed in Chapter 4.

[13] The distinction between indirect and systemic discrimination is quite subtle, and quite academic since the Supreme Court's decision in British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union (1999). The distinction will be discussed in detail in Chapter 4; but for practical purposes the terms can generally be considered to be synonymous.

for more research, especially with respect to the recruitment of women[14] (Linden & Minch, 1994; Lebeuf & McLean, 1997; Policing in British Columbia Commission of Inquiry, 1994; Walker, 1993) and minorities (Jain, 1987; Nelson, 1992). This study, by focussing on sex discrimination at the Police Academy assessment center, attempts to answer the call for more research in police recruitment (e.g., see Burbeck & Furnham, 1985, p. 58) and at the same time attempts to address a "blank spot"[15] in personnel selection literature.

Sex Discrimination

The first female police officers in Canada were hired by the Vancouver and Edmonton police departments in 1912, although they were limited to prison matron duties (Duchesneau, 1997, p. 170; Moore, 1997, p. 38; Polowek, 1996, p. 155). Over the next 30 years, the female police role evolved to include more policing duties, but females were often relegated to separate divisions or bureaus and were restricted to social work (e.g., working with juveniles and female offenders) and community relations. However, in the United States, the Indianapolis Police Department broke new ground in 1968 by assigning females to routine patrol duty (the heart of policing); and in Canada, the Vancouver Police Department followed suit in 1973 (Polowek, p. 156; see also Prindiville, 1975; and Segrave, 1995).

In the late1970s, police organizations began to revise those entry requirements that had an adverse effect against females, in that height and weight restrictions were

[14] Because this research has a legal orientation, gender is defined strictly in terms of being male or female (i.e., sex) rather than in terms of being man or woman, to avoid theoretical discussions that are beyond the scope of this study.
[15] This term was coined by Wagner (1993).

either modified or eliminated in both Canada and the United States. Shortly thereafter, in the mid-1980s, police organizations began to implement recruitment initiatives that targeted females.[16] This was around the time that females achieved integration into specialty sections, such as dog squads, emergency response teams, and major crime sections (Moore, 1997, p. 41; Polowek, 1996, p. 195; see also Segrave, 1995; and Statistics Canada, 1999). And by the mid-1990s, Canadian police departments began to offer increased benefits related to maternity leave and job sharing (Polowek, 1996, p. 138).

Consistent with these developments, since the mid-1970s, when females accounted for less than 1% of police officers in Canada, the proportion of female police officers has increased steadily and consistently. By 1999 females accounted for approximately 13% of all police officers in Canada and 17% in British Columbia (Statistics Canada, 1999, pp. 4, 12-13; see also Women in Policing, Police Chief, 1998, p. 40).[17] However, despite this trend, recent studies in the United States suggest that female applicants have leveled off at approximately 20% (e.g., Dantzker & Kubin, 1998, p. 19; see also Harris, 1999, p. 18).[18] The reasons for such disproportionate applications requires further investigation that is beyond the scope of this study, but the data suggest that a contributing factor is that of systemic discrimination against females entering policing (e.g., see Landrine & Klonoff, 1997, p. ix).

---

[16] Except for a brief period during World War II, when females were actively recruited.

[17] According to Polisar and Milgram (1998), in the United States women accounted for 9.5% of police officers in 1995 (p. 42; see also Dantzker & Kubin, 1998, p. 19; Harris, 1999, p. 18; and Miller, 1998, p. 162), and 11.6% in 1997 (Rank Objections, Law Enforcement News, 1998). In some large cities in the United States, however, women account for over 25% of police officers (e.g., Miami Beach and Madison) (rf. Bureau of Justice Statistics, 1995).

[18] Recruitment data on a provincial or national level are unavailable in Canada, but this leveling of female applicants was also found in this study (see Chapter 5, Table 5-7).

Moreover, females are leaving policing at a disproportionate rate when compared to males (Polowek, 1996, pp. 1, 34; see also Dantzker & Kubin; Felkenes & Unsinger, 1992; Harris; Linden, 1983; and Seagram & Stark-Adamec, 1992). After reviewing the results of various studies in Canada, Polowek (1996) found that the attrition rate for females was sometimes as high as three times the rate for males, although Polowek's data indicate that attrition rates for females have declined since the early 1990s (pp. 18, 34; see also Walker, 1993). Like males, females may leave policing because of dissatisfaction with organizational issues, such as lateral mobility, promotional opportunities, and management style (Dantzker & Kubin, 1998; see also Dantzker, 1994).[19] But, unlike males, females report that they leave policing for reasons related to family (such as the desire to raise children) and for reasons related to sexism (such as harassment, stereotyping, and tokenism) (Polowek, pp. 19, 21-22, 30-34, 111). However, empirical research on the recruitment and retention of females in Canadian (and American) police organizations is limited, making conclusions tentative at best (Moore, 1997, p. 37).

One other institutional hurdle for females is found in the traditional hierarchy of police organizations, where females are proportionately underrepresented in ranks above that of constable (i.e., supervision and management). For example, although females accounted for approximately 13% of all police officers in Canada in 1999, they only accounted for 4.7% of non-commissioned ranks and 2.8% of commissioned ranks (Statistics Canada, 1999, p. 13; see also Women Climbing Police Ranks Slowly, 1998,

[19] Dantzker specializes in the study of police organizations. Because employee job satisfaction may explain everything from attrition to productivity, it is a variable that receives a great deal of attention in organizational theory.

The Globe and Mail).[20] Although underrepresentation of women may be due in part to traditional service requirements (generally the first level of promotion does not occur until after 10 years of service), exacerbated by high attrition rates (Statistics Canada, 1999, p. 12), Polowek (1996) suggests that gender and race may also adversely affect promotions (pp. 200-203; see also Force Has Race Woes, 1996, The Vancouver Sun; McLean, 1997, p. 178; and Women Still Penalized in Promotions, The Vancouver Sun). Although Polowek did not cite any supporting studies, in the United States Martin (1989 & 1991) argues that underrepresentation of women in supervisory ranks is the result of "tenacious resistance to hiring women by high-ranking administrators and officers in many police departments" (cited by Dantzker & Kubin, 1998, pp. 20-21; see also Miller, 1998). This argument is supported in part by the results of a recent national survey conducted by the International Association of Chiefs of Police, which indicate that gender bias may be a factor in nine percent of all promotions (Women in Policing, Police Chief, 1998, p. 36).

As problems of bias and attrition are addressed, it may be expected that over time the proportion of females in management and supervisory positions will likely reflect the total proportion of females in policing. But despite aggressive recruiting efforts, unless other changes are made, females may never be equally represented in policing due to female applications leveling off at approximately 20% of the candidate pool. Here, as noted by Dantzker and Kubin (1998), the question is why does it appear that females are less likely than males to choose policing as a career (p. 29). Sex discrimination, although

[20] Citing Martin (1989 & 1991), Dantzker and Kubin (1998) report similar statistics in the United States, but more recently, it has been reported that females hold 7.4% of executive positions and 8.8% of supervisory positions (Rank Objections, Law Enforcement New, 1998, p. 1).

relevant, may not by itself provide a completely satisfactory answer. For example, Dantzker and Kubin suggest that policing may not be very attractive to females in general, although they found that female police officers reported levels of job satisfaction that were similar to their male counterparts (p. 29). Such research questions are indicative of the "blank spots" in police selection literature, yet, unfortunately, are beyond the scope of the present study and are left for future research.

The Problem of Proof

Because the general focus of this study is on discrimination in the employment context, it has a very practical orientation and so must necessarily turn to the law for guidance.[21] In personnel selection, discrimination will generally occur when an employer or an employment agency (e.g., Police Academy assessment center) treats an individual differently on the basis of personal characteristics or association with a particular group rather than on the basis of capability or merit (Andrews v. Law Society of B.C., 1989, pp. 174-175). For example, making distinctions between individuals or groups based on race, colour, sex, religion, etc., whether intentional or not, will be found discriminatory and in violation of human rights legislation in the absence of a statutory exemption. Although this explanation of discrimination appears rather straightforward, proving that an unlawful "distinction" has occurred is fraught with difficulties and is a matter of intense policy debate (e.g., see Abella, 1984; and Landrine & Klonoff, 1997, Ch. 9).

[21] As a result, this study did not pursue a broad theoretical discussion on issues of justice and fairness. Despite this limitation, the legal approach is based on the moral principle of human worth and equality, regardless of race, religion, sex, etc.

As a result, further compounding the problem of organizational complacency and the problem of identification (including "blank spots" in the literature) is the additional problem of proving systemic discrimination before a court or tribunal (Zinn & Brethour, 1996, p. 1:8). Evidence of systemic discrimination can be qualitative, quantitative, or both (Lasani v. Ontario (Ministry of Community and Social Services), 1993, para. 49, citing Canadian National Railway Co. v. Canada (Canadian Human Rights Commission), 1987),[22] but because systemic discrimination is often revealed in patterns (or effects), a quantitative analysis can be particularly helpful.

Here, statistical analyses[23] can be useful to introduce (rebuttable) evidence of discrimination (e.g., in terms of "significant" differences or relationships), which has been noted by many courts in the United States (U.S.) (Vining, McPhillips & Boardman, 1986, p. 622). For example, in Teamsters v. United States (1977), one of the first important U.S. cases on discrimination, the Supreme Court noted that "statistical analyses have served and will continue to serve an important role in cases in which the existence of discrimination is a disputed issue" (cited in Paetzold & Willborn, 1999, p. xi.). However, in Canada, although statistical analyses are likely to become an appealing alternative to traditional arguments that often rely on evidence of exclusion (e.g., "ratio of non-representation") and simple proportions (Brook, 1990, p. 132; Vizkelety, 1987, pp. 168, 173), to date their probative value is not yet well defined.[24]

---

[22] See also Sheppard (1993, p. 11). For a U.S. example, see McDonnel-Douglas Corp. v. Green (1973).
[23] Statistical analyses here refer to statistical tests of "significance."
[24] For example, in a recent case heard by the Supreme Court (British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union, 1999), the quantitative evidence was limited to simple rates, ratios and proportions.

The Assessment Model

According to Frost (1997), an analysis of sex-based discrimination, which she describes as "Gender Equality Analysis," should assess the differential effect of policies and standards, whether intentional or not, on both females and males (p. 74).[25] By operationalizing "differential effect," a discrimination analysis may indicate systemic discrimination, and such analyses should develop "new approaches and analytical models" (p. 74), although there is no guarantee that any particular model can adequately identify all relevant factors that may contribute to discrimination. With this in mind, the model proposed by this study integrates the U.S. focus on quantitative analysis into the Canadian legal model,[26] which in addition to assessing discrimination in assessment centers may also be useful in other settings. A qualitative analysis, which is included in the proposed model, is also essential because it may identify contextual factors contributing to discrimination that are not amenable to a quantitative analysis.[27]. But because this study intends to highlight the utility of a quantitative analysis in cases of

---

[25] In the report by the Status of Women Canada (1996), it is argued that policy analysis "is incomplete if the impact of gender has not been considered" (pp. 1, 5).

[26] The Canadian and U.S. discrimination models are conceptually similar, which will be apparent if Chapter 4 ("Discrimination") is compared with Paetzold and Willborn's (1999) analysis of U.S. discrimination law. Note also that Abella (1984), in her Report of the Commission of Equality in Employment, cited by the Supreme Court in Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987), recommended an enforcement agency, similar to that found in the United States, to investigate complaints of employment discrimination (i.e., the Equal Employment Opportunity Commission (EEOC)).

[27] This model draws on the U.S. experience with respect to using the quantitative approach in discrimination litigation, which is applied within the Canadian context. It is not suggested that this model is unique; rather, it is an integration of academic research with case law. With respect to academic research, it draws especially on the work (Canadian) of Vining et al. (1986, pp. 676-685) and on the work (U.S.) of Paetzold and Willborn (1999, §§ 2.03, 3.01, 4.01, 4.02, 4.05, 4.08, 5.02). With respect to case law, it is based on the cases (Canadian) of Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987), Lasani v. Ontario (Ministry of Community and Social Services) (1993), and British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union (1999), and on the cases (U.S.) of Hazelwood School Dist. v. United States (1977), McDonnell-Douglas Corp. v. Green (1973), and Teamsters v. United States (1977). Also influential was the Uniform Guidelines on Employee Selection Procedures (1978). See Chapter 4 for a comprehensive analysis of discrimination law in Canada.

systemic discrimination, a qualitative analysis (e.g., legal arguments, anecdotal information, surveys, etc), already well developed in Canadian law, does not receive prominence.

Ironically, within the proposed model, when compared to the test property of reliability and the notion of substantive equality (which focuses on effects), traditional scale validity is found to be relatively unimportant because it is not necessarily a concern legally if discrimination is shown not to exist. If discrimination does exist, it can only be justified by proving the validity of the offending standard (e.g., the assessment center); but validity here is defined in terms of whether the discriminatory standard is necessary to accomplish an employer's legitimate work-related purpose, not in conventional psychometric terms of utility and job performance.

Moreover, conventional demonstrations of validity, usually indicated by way of correlation coefficients, cannot be properly interpreted within the context of substantive equality unless informed by reliability and discrimination assessments. First, reliability ensures that candidates are treated consistently, and is a prerequisite to validity (i.e., correlation coefficients indicating validity are, in part, a function of reliability). Second, although acceptable levels of reliability may exist, a test itself may discriminate. For example, in an assessment center, raters may be consistently scoring females lower than males or blacks lower than whites. Third, should discrimination exist, reliability is an indicator of the extent to which it theoretically exists in the test (i.e., correlation coefficients indicating discrimination are, in part, a function of reliability). And finally, predictive validity (the most common validation technique) is not necessarily threatened by a discriminatory test. For example, males identified as the best candidates in a reliable

yet discriminatory assessment center may do well on validation criteria, such as job performance or promotion, yielding high predictive validity coefficients.[28] As a result, empirical proof of validity falls short of ensuring substantive equality, and so by itself is an inadequate justification of a personnel selection procedure.

The model proposed here, which accommodates either a qualitative approach or a quantitative approach, or a mixed methodology approach, assumes that there are two general phases to a legal analysis of discrimination. The preliminary phase consists of three general overlapping and interrelated steps: (1) identifying the applicable selection procedure (e.g., assessment center); (2) identifying the relevant legal issue (e.g., sex discrimination); and (3) identifying the appropriate groups for comparison (e.g., males and females in the applicant pool). Next, the assessment phase consists of two sequential steps: (1) comparing[29] the groups of interest on the dimension of interest to determine if differences exist; and (2) analyzing any observed differences to determine if they are legally or practically significant. It is in this phase that statistical analyses can be especially helpful in an assessment of systemic discrimination.

Thesis Organization

To address efficiently the multi-dimensional goal of this study (i.e., provide a model to assess discrimination and test its utility by using the Police Academy assessment center as an example), this study is generally organized according to the

[28] For example, Huck and Bray (1976) found that assessment center ratings of black and white women were predictive of both job performance and potential for advancement. However, the scores of white women were significantly higher than those of black women.

[29] Operationally, discrimination must necessarily be defined and measured comparatively, which will be discussed in detail in Chapter 4 (rf. also Status of Women Canada, 1996, p. 3).

structure of the proposed model. In the preliminary phase, Chapter 2 reviews the assessment center method in general, Chapter 3 introduces the theory of reliability and explains its application to assessment centers and to the assessment of discrimination, Chapter 4 provides a comprehensive analysis of discrimination law in Canada, and Chapter 5 describes the Police Academy assessment center, the sample and the method of data analysis.

In the assessment phase, Chapter 6 reports the results of the interrater reliability analysis and the discrimination analysis, which also includes a discussion on whether any observed differences between the scores of males and females were legally or practically significant. Integrating the previous chapters, Chapter 7 summarizes the proposed model and evaluates its utility for assessing discrimination, using the Police Academy as an example. Also, because an important goal of the model is to be proactive, this study makes practical recommendations for improving interrater reliability and preventing discrimination at the Police Academy assessment center.

Of particular importance is that as this study progresses, each chapter theoretically grounds the proposed model. Although this study may appear somewhat onerous to read, it does not suggest that the assessment of discrimination is necessarily onerous. Rather, the goal of this study is to provide a model that allows for an objective and uncomplicated assessment of discrimination in the employment context and in so doing contribute to its prevention. Of course, it must be recognized that discrimination can exist despite the fact that it is not found, which is an inherent limitation of all analytical models.

CHAPTER II: THE ASSESSMENT CENTER METHOD

In the 20th century, the assessment center method was first taken seriously by the German military in World War I, and by the British and American military in World War II. But after World War II, military and government officials seemed to lose interest, especially in the United States (Moses, 1977, p. 9). The private sector, however, adopted the method, which subsequently experienced phenomenal growth on an international scale. Taking World War II as the starting point and North America as the focal point, the purpose of this chapter is to review the literature on the assessment center method. The review will begin with a broad explanation of the method and its theoretical underpinnings. Next, the review will explore the historical roots of the method, tracing its start with the military and its subsequent introduction into the private sector. And last, the review will discuss relevant issues of reliability, discrimination, and validity, followed by an overview of assessment center use by police in the United States and Canada. Although the review is broad (describing the overall framework of the assessment center method), it provides the necessary context for a general discussion of discrimination in assessment centers and for a specific assessment of sex discrimination in the Police Academy assessment center.

An Explanation of the Method

Use of the term "center" to apply to situational testing (i.e., "assessment center") can be traced to the United States military during World War II (Office of Strategic Services (OSS) Assessment Staff, 1948). Although the military was responsible for molding the contemporary assessment center, it was Douglas Bray's "Management

Progress Study" that refined the concept and adapted it for personnel selection by business (Bray, Campbell & Grant, 1974), which is discussed in greater detail in the following section on history.[1] Consistent with Bray's (1999) belief that psychology should focus on behavior, the basic theory upon which the assessment center method rests is that a person's natural behavior is indicative of various abilities and personality characteristics judged essential to success in a particular job or position (see also Bray, 1982, p. 183; and Howard, 1997, pp. 21, 44). Personality characteristics are often loosely categorized as attributes, traits, motives, and attitudes, and include such dimensions as honesty, creativity, resistance to stress, etc. Consequently, assuming that such characteristics are relatively stable, and assuming that candidates behave naturally and spontaneously in situational tests (Greenwood & McNamara, 1967, p. 101, citing Flanagan, 1954), the behavior displayed by candidates in exercises simulating a target position is a reliable and valid predictor of future behavior or performance in that position (Bray, 1982, p. 185; Klimoski & Brickner, 1987, p. 245; Ritchie & Moses, 1983, p. 227).[2] Or as more simply stated in a familiar axiom, past behavior is the best predictor of

[1] The term "assessment center" was formally defined in Standards and Ethical Considerations for Assessment Center Operations in 1975, by the International Congress on the Assessment Center Method (pp. 1-3). In the literature, there is general agreement with this definition, which requires multiple evaluation techniques that must include simulations based on the target position (Reilly, Henry & Smither, 1990, p. 71). According to the 1989 Guidelines and Ethical Considerations for Assessment Center Operations, there is a difference between the terms "assessment center" and "assessment center methodology" (p. 2). Specifically, the term "assessment center" includes all elements identified in the Guidelines as essential to define an assessment center. However, some may use the term "assessment center methodology" to refer to various, but not necessarily all, of these essential elements. Notably, in 1996 the International Congress on the Assessment Center Method voted to change the name of the conference to the International Congress on Assessment Center Methods (Howard, 1997, p. 47). For clarification, in this study any reference to assessment center methodology includes all traditional elements commonly attributed to an assessment center, unless the context indicates otherwise. For a brief historical perspective, see Bray (1999) and Howard (1997). For additional notes on terminology, see Byham (1977a).

[2] Assessment center theory is discussed in more detail in the section on validity.

future behavior, which is an assumption consistently found in assessment center studies (Schmitt, Gooding, Noe & Kirsch, 1984, p. 416).

In spite of this ostensibly simple theory, as noted by Byham, there is difficulty in defining an assessment center (quoted in Mayes, 1997, p. 7). However, there is general agreement in the literature that, basically, an "assessment center" is a standardized personnel selection method that involves multiple evaluation techniques, with a particular focus on behavior. Traditionally, according to user surveys, the purpose of an assessment center has been to assess a candidate's overall potential for promotion (46% reported) or employment (22% reported) in an organization (Gaugler, Rosenthal, Thornton & Bentson, 1987, p. 497), but recent research indicates that assessment centers are used equally for promotion and selection (Spychalski, Quinones, Gaugler & Pohley, 1997, pp. 72, 78). Other uses include developmental planning and training and basic research. Although traditional personnel selection methods such as interviews and various psychometric assessment tests may be used, an assessment center's defining characteristic is the objective assessment of a candidate's behavior within the context of a job-related simulation (Bray, 1982, p. 183; Howard, 1997, pp. 15, 44; 1974, p. 117; see also "Assessment Center Defined" in Guidelines and Ethical Considerations for Assessment Center Operations, 1989, p. 2).

The most common simulations, otherwise known as exercises, include in-basket exercises, group discussions, interviews, fact-finding exercises, decision-making exercises, oral presentations, and written communication exercises (Byham, 1999). Exercise content is usually developed through a job analysis and is designed to simulate situations that one would normally encounter in the target position. Despite the practical

necessity of such exercises to provide job-related context, the emphasis is not on candidate performance of the exercise but on candidate behavioral characteristics (commonly known as dimensions) displayed in the exercise.[3] Consistent with the job analysis, the exercises are designed to elicit behavior that will provide examples of a candidate's abilities on the dimensions of interest (e.g., practical intelligence, integrity, flexibility, initiative, etc). These dimensions, equally important in their own right, theoretically underpin the specific skills, knowledge, abilities, values, attitudes and motives that have been identified as critical to success in the target position (Byham, 1999; Howard & Bray, 1988, p. 9).[4]

Consistent with the use of multiple techniques and scenarios in order to achieve a holistic evaluation of a candidate, the assessment center method uses multiple raters. Notably, the raters are not necessarily psychologists but are often employees, selected from the occupation or profession of interest, who are trained to observe, record, and evaluate a candidate's behavior in the exercises. Raters systematically and independently record their observations and subsequently assign a numerical score for each dimension, indicating on a standardized scale the level of competence for the target position. After the exercises are concluded, the raters discuss and defend their observations in an integration session moderated by an administrator. The purpose of this session is for the raters to reach consensus on each candidate's potential for success in the target position by means of an overall score (Guidelines, 1989, p. 2). The typical assessment center

[3] Traditionally, an assessment center focuses on dimensions, but assessment centers have been developed that focus on exercises or tasks. This issue will be discussed later in this chapter.
[4] Some dimensions (or exercises) may be generalized across different organizations (Byham, quoted in Mayes, 1997, p. 3).

(such as the Police Academy assessment center)[5] uses three or four raters for each group of five or six candidates, and generally lasts from one to three days depending upon the instruments used and the complexity of the target position.

Notably, the assessment center is not an evaluation instrument but a procedure or a method (Howard, 1974, p. 116; 1997, p. 47; Norton, 1981, p. 562), the key components of which include a job analysis, behavioral classifications according to job-related dimensions, multiple evaluation techniques, multiple simulations, multiple assessors, standardized scoring, and a data integration session (Guidelines, 1989, pp. 2-4). Because the assessment center is a method, it is easily adaptable to any occupation or profession in any country, which helps explain its use internationally (Byham, 1999; Howard, 1997, p. 11). For example, the dimensions considered critical to policing, and how such dimensions are behaviorally exhibited, may differ across cultures (Briscoe, 1997). However, it is not what dimensions are used and how they are defined that characterize an assessment center; rather it is the method itself that specifies a unique standardized procedure by which such dimensions are identified and assessed.

Although the assessment center method may seem cost prohibitive[6] and cumbersome because it is demanding in terms of time and personnel, it has a number of advantages (which will be reviewed in detail) over traditional personnel selection methods that warrant attention by an employer. First, because the assessment center is a

---

[5] Gaugler et al. (1987) suggest that the "typical" assessment center does not exist, which is true when one considers the variety found in operational centers (p. 494; see also Bender, 1973). Nevertheless, for a good general description of a typical assessment center, see Byham (1980a, pp. 24-25), Cohen (1980, pp. 989-990), and Jaffee (1984, pp. 2-5).

[6] In a recent survey of assessment center practices in the United States, Spychalski et al. (1997) found that the average cost associated with a single candidate was $1,730.00 (U.S.), with a standard deviation of $5,192.00 (U.S.) (p. 82).

method rather than an evaluation instrument, it can be specifically tailored to any organization, within any culture. Moreover, because it is a method and not a place, it offers flexibility in terms of time and location. Second, the assessment center method is able to measure traits or dimensions not adequately measured by written tests, such as interpersonal skills and oral communication abilities. Although an interview may provide an opportunity to evaluate these dimensions, it is usually limited to approximately an hour while the assessment center usually lasts a full day. Third, the assessment center method, compared to traditional written tests, commonly results in less discrimination against protected groups. Fourth, the assessment center method compares well with traditional selection methods (e.g., interviews and written tests) with respect to reliability and validity. And finally, because the assessment center method possesses a high level of face validity (it is easily understood by employers, raters, and candidates), it is well accepted in the employment context.

History

An examination of the historical "roots" of the assessment center provides the context necessary to facilitate a more complete understanding of the assessment center method. According to Adair and Moon (1977) the first documented use of systematic behavioral analysis to assess candidates for organizational selection purposes is found in ancient China, where it was used to identify civil servants. According to these authors, this system was "adopted and modified by the East India Company around 1830 for the selection of colonial administrators" (p. 2, citing Parkinson, 1957). However, according to Tielsch and Whisenand (1977) the first recorded use of behavioral analysis for

selection purposes is found in the Old Testament, which describes how it was used by Israel's military to identify the best candidates for a special military operation.

The Military Tradition

According to the Biblical account found in the book of Judges,[7] Gideon, a judge and military leader of the nation of Israel, was leading his men to battle, when God said, "Bring them down unto the water, and I will test them for thee there.... And the number of them who lapped, putting their hand to their mouth, were three hundred men; but all the rest of the people bowed down upon their knees to drink water. And the Lord said to Gideon, By the three hundred men who lapped will I save you, and deliver the Midianites into thine hand; and let all the other people go every man unto his place" (ch. 7). Therefore, by objectively analyzing the behavior of those soldiers who came to drink water, Gideon was able to select the best men for battle.

Ironically, in the 20th century the military was the first organization to use behavioral simulations to identify personnel for special assignments. According to Yan and Slivinski (1976), the origins of the contemporary assessment center began with the Germany military in 1915 of World War I (cited by DuPerron, 1997, p. 10). This account is not inconsistent with that of MacKinnon (1975a; see also 1975b),[8] who argues that contemporary assessment centers owe significant credit to the pioneering work of

[7] King James version.

[8] MacKinnon, a "pioneer" of the assessment center, was also described by Moses and Byham (1977) as an "able historian" (p. 13). MacKinnon was Director of the original OSS assessment center at Station S during World War II, a contributing author to the text Assessment of Men (Office of Strategic Services (OSS) Assessment Staff, 1948), and director of the Personality Assessment Institute, Berkeley, University of California, for 20 years.

German military psychologists in the 1930's, most notably M. Simoneit (p. 1; see also Moses, 1977, p. 8).

Before World War II, both the British and American military used traditional selection criteria (e.g., education) and techniques (e.g., written tests, interviews, and background investigations) to select personnel for officer positions. In England, by 1941, it had become obvious to the War Office Selection Boards (WOSB) that the traditional selection methods were not working due to alarming failure rates at Officer Cadet Training Units and the number of officers who had mental breakdowns (Tielsch & Whisenand, 1977, p. 10). In the United States, the President and Congress established the Office of Strategic Services (OSS) during World War II for a variety of duties (Girodo, 1997, p. 239). These duties included establishing a network of international agents to gather strategic information from countries considered a national threat. Another duty was to send agents into enemy territory to destroy military targets, aid resistance groups, and distribute propaganda (Office of Strategic Services (OSS) Assessment Staff, 1948).

In an attempt to improve upon conventional methods for selecting these agents, in the early 1940s OSS psychologists, led by Henry Murray[9] and his Harvard University colleagues, notably Donald MacKinnon, conducted some innovative research in assessment techniques.[10] This research was unique in psychology because it attempted to understand normal personality holistically—an "organismic" approach vis-a-vis the

[9] In an interview with Mayes (1997), Bray stated that previous research by Murray (e.g., Explorations in Personality, published in 1938) influenced the design of the OSS Assessment Centers (p. 1; see also Bray, 1982, p. 180; and MacKinnon, 1975a, p. 1).

[10] The OSS published a book in 1948, entitled "Assessment of Men," which was the first formal publication of the assessment center method in the United States.

traditional "elementalistic" approach (Bray, 1982, p. 181; Bray, 1999; Howard, 1974, p. 117).[11] This research was also unique in personnel selection because it used interactive exercises and behavioral simulations in addition to traditional interviews and written tests, as suggested by the WOSB in Great Britain (Bray, 1982, p. 180). In the United States, the OSS is generally credited with establishing the first assessment center used for personnel selection.

The OSS assessment center methodology included the following: conducting a job analysis on the target position; identifying behavioral attributes considered necessary for success; constructing scales on which raters would assess candidates on each key attribute, including an integration procedure in which overall ratings and recommendations would be decided; creating assessment exercises that simulated field conditions; and finally, designing the format for a personal history interview (Moses, 1977, p. 9; Tielsch & Whisenand, 1977, pp. 11-12). At the same time, the British WOSB and the British Civil Service Board were conducting similar research for military and civil service officer selection (MacKinnon, 1975a, p. 1; Moses, 1977, p. 9).

A classic example of a situational exercise where raters observe and record candidate behaviors is the famous OSS "Brook Test" (OSS Assessment Staff, 1948, pp.

[11] Although personality is obviously a factor in a person's behavior, an operational assessment center does not claim to assess personality. Rather, it assesses characteristics or attributes (i.e., dimensions), which are defined generally, through behaviors that have been judged necessary for success in the target job. In other words, it is a candidate's ability to perform a particular job that is being assessed, not the candidate's personality. Personality characteristics are therefore not analyzed or diagnosed in any psychological sense (and they do not tell the whole story about a candidate's ability to successfully perform a particular job). Rather, defined behaviorally, specific attributes (e.g., the ability to delegate, communicate clearly, and organize others) are traditionally clustered into general dimensions (e.g., leadership), which are measured against objective standards. As noted by Bray (1982), personality characteristics "are, in fact, often interdicted because of possible legal problems concerning predictive validity and because it is feared that including them will dilute attention to behavior, the hallmark of the assessment center method" (p. 183). In summary, the "organismic" approach of the assessment center emphasizes the observation of behavior for measuring performance on a particular job-related dimension.

95-97), ironically somewhat similar to Gideon's water test. Here, a group of candidates (approximately six) was taken to an isolated, natural area, which included trees and a shallow, narrow, quiet brook. The banks were approximately eight feet apart, and on one side was a log and on the other a heavy rock. On the side where the candidates were taken, where the log was located, boards were scattered about (none long enough to span the brook), along with three lengths of rope, a pulley, and a barrel without ends.

The raters instructed candidates to imagine that the brook was a deep, raging river that required them to work from the tops of sheer banks. Their task was to take sensitive equipment (camouflaged as a log) to the opposite bank and return with explosives (camouflaged as a heavy rock). Candidates were given ten minutes to plan and then instructed to begin the exercise. During the exercise (known as a "leaderless exercise") raters (usually three) would record all behaviors and conversations of candidates to whom they were assigned, which would theoretically yield information on dimensions such as leadership, applied intelligence, initiative, energy, interpersonal skills, and physical ability (Girodo, 1997, p. 240; MacKinnon, 1977, pp. 20-21; Tielsch & Whisenand, 1977, p. 14). This is an illustrative example of the heart of the assessment center method.

After World War II, interest in the assessment center waned. The military and government did not pursue the assessment center method (apart from limited use by the United States Central Intelligence Agency (CIA, a product of the OSS), and the British Civil Service). However, private industry adopted it, improved it, and watched it grow into an international phenomenon (Hinrichs, 1978, p. 596; Howard, 1997, p. 17; MacKinnon, 1975a, p. 2; Maher, 1984, p. 20; Moses, 1977, p. 9; Tielsch & Whisenand, 1977, p. 24). Although the purpose would have been the same for industry, military, and

government (i.e., efficiently identifying the best candidates for the target job), industry obviously had greater faith in the method's potential.[12] While the exercises and dimensions obviously varied from those used by the military, the method was essentially the same (Moses, p. 9).

Pioneers in the Private Sector

There is little doubt that the American Telephone and Telegraph Company (AT & T) was a leader in the private sector in the area of personnel selection research, initiating a unique and unprecedented study that began in 1956 under the direction of Douglas W. Bray. In a presentation at the 26th International Congress on Assessment Center Methods in Pittsburgh, Pennsylvania, Bray (personal communication, May 11-14, 1998) described how he became involved in the AT & T project (see also Mayes, 1997). After reading the "Assessment of Men" (OSS, 1948), while nearing the completion of his doctorate in psychology at Yale, Bray became fascinated by the assessment center method, but had no opportunity to explore this unique personnel selection method.[13]

AT & T had long been interested in the changes of personal characteristics of managers as they progressed through their careers, which evolved into an interest in the relationship between college education and success in management and on validating methods for managerial selection (Bray, 1982, p. 182; Howard & Bray, 1988, p. ix). Although in the early 1950s AT & T was committed to a long-term study, it was unsure of an appropriate methodology. In 1956 Bray came to the attention of AT & T and

[12] MacKinnon (1975a), citing Loacher, 1974, reports that the education and policing professions did not show interest in assessment centers until the early 1970s.

[13] Bray's research on observational measures of performance for the Aviation Psychology Program in the Army Air Forces, during the last part of W.W. II, was included in the OSS volume (Bray, 1999).

Michigan Bell (a subsidiary at the time of AT & T)[14] through a personal reference. Being impressed by Bray, AT & T offered him a job along with a research opportunity on a "silver platter," which was to conduct an extended longitudinal study on managers and their careers. Bray accepted, seizing on the opportunity to test the assessment center theory he found so intriguing in the OSS publication. After persuading Robert K. Greenleaf (AT & T's then director of personnel research) of the potential benefits, Bray began work on what would become a lifelong project, which he named "Management Progress Study" (Bray, personal communication, 1998, supra).

Even before any results of Bray's research became available, the assessment center method gained attention because of employer dissatisfaction with traditional personnel selection methods. For example, according to Bray, in 1958 Michigan Bell was dissatisfied with their ability to select foremen and invited him to implement an operational assessment center. Similarly, in 1961, Dartmouth College in Atlanta complained that businesses were sending "dunces" to their executive development courses (Bray, personal communication, 1998, supra).

In the mid-1960s, William C. Byham, who was employed by J.C. Penney, "galvanized the assessment center movement," initiating the first assessment conference in 1969. In 1970 he published his classic study, "Assessment Centers for Spotting Future Managers," in the Harvard Business Review (Bray, personal communication, 1998, supra; Mayes, 1997, p. 8). Also in 1970, Byham conceived of Development Dimensions

[14] On January 1, 1984, AT & T was subject to divestiture, which split 23 Bell System telephone companies, such as Michigan, Chesapeake & Potomac, New York, Pennsylvania, Northwestern, and Mountain, from parent ownership (Howard & Bray, 1988, pp. 3, 6).

International (DDI), which markets assessment center materials and provides consulting services, and invited Bray to be a co-founder of the new company (Mayes, p. 8).

In 1971, while Byham and Bray were at an American Psychological Association symposium, Bray met Ann Howard, whom he later married. In 1971 and 1972, DDI marketed the assessment center method on an international scale, in countries such as South Africa, Brazil, and Japan. In 1973, DDI initiated the first International Congress on the Assessment Center Method in Virginia, and one of the first major discrimination studies on the assessment center method was conducted (Huck, 1974). In 1974 Howard published what was to become a "classic article" on assessment centers, entitled "An Assessment of Assessment Centers," and in1975 she finished her doctorate and joined Bray's research team at AT & T (Bray, 1999), where she and Bray directed the Management Progress Study and subsequent assessment center studies.

Bray is acknowledged as the father of the assessment center method (e.g., see Byham, personal communication, May 12, 1998; Howard, in Howard & Bray, 1988, p. 3; MacKinnon, 1975a, p. 5; and Lee, 1985, p. 69). Without question, both Bray and Byham are the pioneers of the contemporary assessment center method. Moreover, they are the most influential writers in the field—their research is cited in almost every journal or magazine article written about the assessment center method since 1960. Bray has retired from AT & T after 28 years as a senior researcher, but continues to serve as chair of DDI, while Byham continues to serve as president and chief executive officer of DDI.

The Management Progress Study

The Management Progress Study[15] involved assessing novice managers from 1956 to 1960 and tracking them over the next few years, up to a maximum of eight (Bray et al., 1974; see also Bray, 1964; and Bray & Grant, 1966). The study was unique not only because it documented the first use of the assessment center method vis-a-vis traditional psychometric methods in business (Bray & Campbell, 1968, p. 36), but also because of its longitudinal methodology and duration (eight years) within a single organization, and the size and comprehensiveness of its data collection. MPS by itself established the assessment center as a valid method in the field of personnel selection (MacKinnon, 1975a, p. 2; Moses, 1977, p. 10). In the early 1970s MPS became known as "The Study" in assessment center methodology (Howard, 1974, p. 122; see also Hinrichs, 1978, p. 596) and today remains the classical model for assessment centers (Klimoski & Brickner, 1987, p. 255; see also Lowry, 1997, p. 53) and continues to be the most widely known and cited study on the subject.

Systematically documenting the growth and development of 422 new AT & T first level managers in six of the 23 Bell System companies, one purpose of MPS was to determine if the assessment center method could predict those managers who would have successful careers (Howard & Bray, 1988, p. 4; Moses, 1977, pp. 9-10). The sample consisted of all 274 college graduates who were laterally hired into first level

[15] Hereafter referred to as MPS. The original MPS was followed up by a second study, named MPS:8, which after eight years re-assessed the remaining MPS subjects, and a third study, named MPS:20, which after twenty years re-assessed the remaining MPS subjects (Howard & Bray, 1988). Finally, in 1977 Howard and Bray (1988) launched a new but complementary study, named the Management Continuity Study (MCS), with new subjects that this time included women and minorities (pp. x-xi). Bray was the director of the AT & T MPS and MCS studies until his retirement in 1983, at which time, Howard, the associate director, succeeded him as director (Howard & Bray, p. xiii).

management and 148 randomly selected employees who had no college education but were promoted from within. A limitation of this particular sample was that it consisted of only white males, although other MPS samples included women and minorities. This limitation is not surprising, however, considering the historical context and company goal to select those participants who would have a reasonable chance of reaching middle management levels or higher (Bray, 1982, p. 188; Howard & Bray, 1988, pp. 5-6).

MPS involved putting managers through a three and one-half day assessment center. Each year data were collected on the managers to track their career progress. After twenty years, of the college graduates who were predicted by MPS to reach the fourth level of management (out of seven levels), 60% did so, while only 24% who were assessed less favorably did so. Similarly, of the sample without a college education who were predicted by MPS to reach the third level, 58% did so, while only 22% who were less favorably assessed did so (Howard & Bray, 1988, p. 9; see also Bray, 1982, p. 185; Huck, 1977, p. 266; Lee, 1985; Moses, 1977, p. 10). Notably, the data were uncontaminated—neither candidates nor company management were aware of assessment center results on any individual.

Even though the results of MPS were unknown, in 1958 the Michigan Bell Telephone Company established the first operational assessment center in the business world (Bray & Campbell, 1968, p. 36; MacKinnon, 1975a, pp. 3, 5). Influenced by MPS, and following the example of Bell, other companies began using the assessment center method; for example, Standard Oil, International Business Machines (IBM), General Electric, Sears, Caterpillar Tractor, and even the Canadian government (Byham, 1977a, p. 40; 1980, p. 24; MacKinnon, p. 3; Moses, 1977, pp. 10-11). As a result, the method

was applied to diverse occupational groups, such as upper management, sales personnel, and engineers, and eventually to professions such as education and policing. Today, according to Byham's rough guess, approximately "80 percent of the Fortune 500 companies use assessment centers somewhere in their organizations" (quoted in Mayes, 1997, p. 7).

The motivation for pursuing the assessment center method, as previously stated, was to identify in the most efficient manner the best candidates for the target job. For business, this is translated into cost benefits, which generates organizational interest "in utilizing any method of selecting managers that promises to be more efficient than traditional ones" (MacKinnon, 1975a, p. 4). Another attraction was its basic simplicity. As stated by MacKinnon, psychologists offer similar services, which for the layperson, however, are cloaked in "an aura of mystique," while assessment centers are "less mysterious, more open, and—above all—have high face validity" (pp. 4, 28; see also Gavin & Hamilton, 1975, p. 176; Jaffee, Cohen & Cherry, 1972, p. 26; Lowry, 1996, p. 307; and Stinchcomb, 1985).

Even though issues such as job analysis, reliability, and validity may involve sophisticated research designs and statistical techniques, the assessment center concept is as easily understood today as it was for Gideon, or the raters at the first "Brook Test." This is because the method is job-related, based on the assumption that if candidates can exhibit appropriate behavior in relevant simulations, they will perform similarly in the target position. As MacKinnon (1975a) stated, "To many, this seems both plausible and fair, and is one of the reasons why assessment centers have won such widespread acceptance (p. 4).

Reliability

Kulis (1987) suggests that the high reliability coefficients reported for assessment centers are "part of the folklore of the field" (p. 128). Although there is some support for this (e.g., Hinrichs &Haanpera, 1976), most research indicates more than acceptable levels of reliability. MacKinnon (1975a) found that the research supported "quite clearly the conclusion that interrater reliabilities in assessment evaluations are sufficiently high to justify their further use" (p. 14), as did Huck (1977) who found the evidence "rather conclusive" (p. 277). For example, reliability coefficients between .60 and .90 are commonly reported even though raters are required to make judgments in a variety of assessment center configurations. Specifically, coefficients greater than .60 are usually reported when raters classify behaviors (Gaugler, 1987), rate overall performance in an exercise (Bray & Grant, 1966; Greenwood & McNamara, 1967), rate dimensions within an exercise (Borman, 1982), rate dimensions across exercises (Schmitt, 1977), or make judgments on overall scores (McConnell & Parker, 1972; Sackett & Hakel, 1979; Schmitt, 1977).

Interestingly, research has shown that the ratings of properly trained lay assessors can be as reliable as those of professional psychologists (Bray & Campbell, 1968; Greenwood & McNamara, 1967, pp. 101, 105-106; see also MacKinnon, 1975a, p. 14; and Huck, 1977, p. 278). On the other hand, Gaugler et al. (1987), in a meta-analysis of assessment centers, reported that psychologists provided more valid ratings than managers did (see also Lowry, 1993, p. 489). Such contradictory findings, though, may be due to interrater reliability being more a function of assessor training than of expertise in psychology (Hinrichs & Haanpera, 1976, p. 39; Huck, p. 279; Kulis, 1987, pp. 130-

136). As Byham (1977b) noted, "There are striking differences in various assessment [center] programs in the amount of training given assessors" (p. 89), a variable that has not been controlled in these studies.

Given that interrater reliability coefficients of .60 in assessment centers are generally considered acceptable (Kulis, 1987, p. 128),[16] the research suggests that reaching an acceptable level is not a problem. Rather, the problem is with how interrater reliability is reported in the literature—correlation and analysis of variance methods are common, but their specific application or technique is generally not reported (Fleenor, Fleenor & Grossnickle, 1996, pp. 368, 379). For example, Shechtman (1992) reported interrater correlations between .66 to .80, which he described as "high" (p. 384), but did not report the technique. Tinsley and Weiss (1975), in what may be somewhat of an overstatement, describe this lack of detail as an unacceptable practice (p. 359).[17]

However, in a recent survey of assessment center practices Spychalski et al. (1997) found a practice worse than inadequate reporting, which was that 21.4 % of respondents reported that they did not even evaluate reliability (p. 80). Considering that reliability results may be used to assess discrimination, and that it is a pre-condition to validity, this is probably more properly judged as an "unacceptable" practice. For those assessment centers that did evaluate reliability, Spychalski et al. unfortunately did not report average coefficients, although they did report that "interassessor agreement was the most popular" (p. 79).

---

[16] The minimum standard for interrater reliability at assessment centers will be discussed in detail in Chapter 3.

[17] In an example of what should be reported, in Berry v. Omaha Mendenhall (1992) described how Byham had used correlational analysis (i.e., Spearman's "rho") to obtain a coefficient of .84 (p. 62).

Discrimination

Discrimination in personnel selection has been especially topical since the mid-1960s (Gavin & Hamilton, 1975, p. 170), which Jaffee (1984) suggests was a factor accounting for the rising popularity of the assessment center method (p. 1). As of 1980, although several civil charges of discrimination were filed against assessment centers in the United States,[18] according to Byham (1980b) none were successful (p. 30). This is especially noteworthy in light of the Griggs v. Duke Power Co. (1971) case. In a 1980 U.S. case involving allegations of discrimination (Firefighters Institute for Racial Equality v. City of St. Louis, 1977), the court stated, "In our view the assessment center portion of the examination comes the closest to comporting with the [Equal Employment Opportunity Commission] Guidelines and would thus be the fairest basis for the selection of the eight black firefighters" (Byham, 1980c, p. 5; see also Hurley, 1987, pp. 38-39). In a recent interview, Byham stated that assessment centers in general "have been found to be more fair to protected groups than any other selection methodology" (Mayes, 1997, p. 12), a conclusion that appears to be generally supported by the research (Bobrow & Leonards, 1997; Fitzgerald & Quaintance, 1982; Thornton & Byham, 1982), and the American courts (Byham, 1980; Frank & Preston, 1982; Hurley, 1987; Kulis, 1987).

In as early as 1972 Jaffee et al. reported that no discrimination between black and white candidates occurred at a supervisory assessment center. However, in 1976 Huck and Bray found a discriminatory effect in a study comparing the performance of black women to white women in a supervisory assessment center at Michigan Bell.[19] Although

[18] In Canada, there are no published legal challenges of the assessment center.
[19] This study was based on Huck's (1974) doctoral research.

the predictive validity coefficients of black women were similar to those of white women, in most analyses the scores of white women were significantly higher than those of black women.

Despite some evidence of racial discrimination, no sex discrimination was found at Michigan Bell's management assessment center. Moses and Boehm (1975) compared the performance of women assessed between 1963 and 1971 with a comparable sample of men who were assessed separately (Moses, 1972). Here, Moses and Boehm found that the distribution of scores for females was nearly identical to that of men, concluding that "the assessment-center method appears to be a logical means for providing equal opportunity to women for promotion into management positions and advancement within managerial levels" (p. 529).

In another study at Michigan Bell, Ritchie and Moses (1983) compared the performance of women, who were originally assessed between 1973 and 1974 (Bray, 1976), with that of men assessed in Bray's MPS. Although Ritchie and Moses reported that the proportion of women who were assessed as having middle management potential (42%) was similar to that of men (40%), they did not compare the average scores of women with men. However, rather than looking for sex discrimination per se, the authors were exploring the theory that the management style of successful women managers was similar to that of men.[20] Ritchie and Moses concluded:

> The current study reinforces the principle that successful women managers are quite similar to their male peers and that their management potential is predictable using the same techniques. It appears increasingly clear that differences in

[20] For an exploration of this topic within the context of assessment centers, see "Men and Women in Management" in Howard and Bray (1988, pp. 266-302).

management potential are far more attributable to individual rather than sex differences. (p. 231)

Similarly, in the Management Continuity Study (MCS), a parallel study to MPS but conducted twenty years later (1977-1982), the proportion of women who were predicted to reach middle management potential (53%) or beyond (13%) was similar to that of men (51% and 11% respectively) (Howard & Bray, 1988, p. 296).

In 1980 Parker reported the results of a study that surveyed 3,365 participants from 58 organizations (mostly manufacturing). Discrimination was assessed by measuring the relationship (correlation) between overall performance scores at an assessment center and personal characteristics, including sex, race, age, and education. Results indicated that overall assessment scores were generally unrelated to age and sex, weakly but significantly (up to .15) related to race, and strongly related to education (up to .32). Although the correlation on race was statistically significant, according to Parker the results were not of practical significance because differences in performance were small (p. 67).

In a more recent study, Hoffman and Thornton (1997) examined the utility of the assessment center method and traditional selection procedures with respect to cost and discrimination. From a utility perspective, this study is different because it argues that validity, if not balanced by a consideration of discrimination, is insufficient in personnel selection research (Hoffman & Thornton, p. 565; see also Cronbach, Yalow & Schaeffer, 1980, p. 693; and Norton, 1981, p. 565). For example, although cognitive ability[21] tests

[21] In this study, cognitive ability is used synonymously with mental ability, which is discussed in more detail in Chapter 5.

are cost effective and have high predictive validity, especially for complex jobs (Ree & Earles, 1992; see also Howard, 1997, p. 19), including police work (Gavin & Hamilton, 1975, p. 172), they often produce adverse effect against minorities (Gottfredson, 1988; Howard & Bray, 1988, pp. 339-342; Hurley, 1987, p. 25; Singer, 1993, p. 23). On the other hand, although assessment centers are more expensive with less predictive power, most evidence generally indicates that they may result in reduced levels of sex and race discrimination.

Using the Uniform Guidelines' (1978) four-fifths rule (rf. Equal Employment Opportunity Commission, 1978),[22] Hoffman and Thornton (1997) did not find discrimination until the 60th percentile in the assessment center but as early as the 20th percentile in cognitive ability examinations. Subsequently, using the Brogden-Cronbach-Gleser gross utility model, Hoffman and Thornton found that the assessment center produced higher utility than cognitive ability examinations "when cut scores on each are set so as to eliminate adverse impact, even though the AC has slightly lower validity and costs considerably more" (p. 464).

Researchers have also investigated the interaction between raters and candidates. An important study is that of Walsh, Weinberg and Fairfield (1987) because it is one of the few studies to report discrimination in an assessment center, and only one of two known studies to investigate interaction effects between rater sex and candidate sex in an assessment center. No only did Walsh et al. find a significant main effect favoring women over men, but the authors also found a significant interaction effect between rater sex and candidate sex that favored women over men (p. 305). Here, female candidates

[22] The four-fifths rule is discussed in greater detail in Chapter 4, and its computation in Chapter 5.

were rated higher than male candidates when an all male assessor group rated female candidates, but not when the assessor group included males and females (pp. 306-307). Walsh et al. found the direction of the bias to be "surprising," although it was consistent with some previous research that found a main effect in favor of women (rf. Friedman, 1984, p. 13; Hamner, Kim, Baird & Bigoness, 1974; Mobley, 1982; and Peters, O'Conner, Weekley, Pooyan, Frank & Erenkantz, 1984).

On the other hand, Lowry (1993), after examining the effects of rater age, gender, race and education on overall candidate scores, found no evidence of discrimination (p. 497; see also Friedman, 1984; and Murray, 1996). Similarly, in the most recent published study on sex discrimination in assessment centers, and the only other study to investigate interaction effects between rater sex and candidate sex in an assessment center, Shore, Tashchian, and Adams (1997) found no evidence of discrimination. Contrary to Walsh et al. (1987), Shore et al. found no significant differences between the scores of men and women candidates on any of the assessment dimensions in any of the exercises (main effects), nor did they find significant interaction effects between rater sex and candidate sex. Consequently, they concluded that their findings "attest to the inherent fairness of the assessment center method and add to the growing body of literature showing that actual human resource evaluations are less prone to gender bias than simulated decisions made in laboratory settings" (p. 201).

Published literature, then, is inconclusive. Assessor characteristics (demographic and personality) may explain some of the variance in scores, but the evidence is limited (Bartels & Doverspike, 1997, p. 153; Lowry, 1993, p. 489). Although little is known about the mediating (i.e., intervening) effects of variables such as age, education,

cognitive ability, and ethnicity on candidate scores (Shore et al., 1997, p. 200), evidence does indicate some adverse effect towards disadvantaged racial groups, such as black people (Howard & Bray, 1988, pp. 339-440). Nevertheless, there is little published research on discrimination in assessment centers (Friedman, 1984, 10; Shore et al., p. 192; Singer, 1993, p. 31), and to make matters worse there are limited data on women and minorities participating in assessment centers (Fitzgerald & Quaintance, 1982, pp. 13, 19; Ritchie & Moses, 1983; and Spychalski et al., 1997, p. 72). Of the studies that have been conducted on sex discrimination, rather than finding discrimination against women, some studies report a main effect in favor of women. Unfortunately, no studies have been able to offer a definitive empirical or theoretical explanation for this phenomenon. However, the overall evidence to date indicates that the assessment center method often results in less discrimination in the personnel selection process, especially when compared to traditional written examinations,[23] although it must be recognized that discrimination in the selection process can occur in areas other than testing (e.g., the recruitment of candidates for testing).

Validity

The assessment center method seems to make sense—predicting a candidate's performance in an organization on the basis of observed behavior in job-related simulations has face validity. As MacKinnon (1975a) stated, "To many, this seems both

[23] Although it is beyond the scope of this study to explore why the assessment center method results in less discrimination than traditional pencil and paper procedures, explanations include the emphasis on observing and recording behavior directly related to the target job, multiple raters, training, and a scoring procedure that requires individual raters to defend their scores in an integration session (Hinrichs & Haanpera, 1976, p. 31; Lowry, 1993; Parker, 1980).

plausible and fair, and is one of the reasons why assessment centers have won such widespread acceptance" by employers and employees (p. 4; see also Byham, 1980, p. 24; Howard, 1997, p. 11; Jaffee, 1984, p. 3; and Teel & DuBois, 1983, p. 87), by the courts (Hurley, 1987, p. 42); and by the general public (O'Hara & Love, 1987). However, face validity by itself is insufficient, which is why MPS was so significant in establishing the assessment center as a valid personnel selection method.

The Case for Predictive Validity

A key research question addressed by MPS was that of predictive validity (otherwise known as criterion-related validity), which "is at issue when the purpose is to use an instrument to estimate some important form of behavior that is external to the measuring instrument itself, the latter being referred to as the criterion" (Carmines & Zeller, 1979, p. 17, citing Nunnally, 1978). The empirical measurement indicating the degree of predictive validity is usually the correlation between the predictor (e.g., assessment center) and criterion (e.g., job promotion).

While MPS continues to be cited today in support of validity claims for the assessment center method, other early research studies also confirmed the validity findings of MPS (e.g., Moses, 1973; Hinrichs, 1978; for a review of early studies see Byham, 1980; and Howard, 1974). In follow-up investigations at AT & T, Bray and Campbell (1968) identified candidates with sales skills rather than those with managerial skills. Again, the data were uncontaminated and the predictions were statistically significant, providing additional evidence for the validity of the assessment center method. Moreover, a number of studies have found that the assessment center method is equally valid for both women and men (Moses and Boehm, 1975; Ritchie & Moses,

1983), as is the case for both black women and white women (Huck & Bray, 1976). Despite the limited number of studies, because of their comprehensiveness and quality, by the mid-1970s there was general acceptance of assessment center methodology in business and industry (Friedman, 1984, p. 9; MacKinnon, 1975a, pp. 5, 17).

Predictive validity (as opposed to construct validity) was the method of choice due to the purpose of operational assessment centers—to identify the best candidates for the target job. The most common validation criteria were (and still are) job performance and job progress. Job progress generally yields higher validity coefficients than does job performance (Huck & Bray, 1976), but the highest coefficients are found in basic research studies (Gaugler et al., 1987, p. 503). In MPS, the validation criterion was job progress (i.e., promotion). Here, predictive validity was confirmed by correlational analysis, the results yielding a coefficient of .44 (MacKinnon, 1975a, p. 17). According to MacKinnon, similar results were obtained in other studies, which reported coefficients ranging from .29 to .63 (p. 17; see also Howard, 1974, p. 124).

In an earlier study at IBM, Wollowick and McNamara (1969) reported an average predictive validity coefficient of .37, which has proven to be the standard in two subsequent meta-analyses. In 1984 Schmitt et al. found an average validity coefficient of .41, and in 1987 Gaugler et al., who analyzed fifty assessment center studies, found a average validity coefficient of .37 (see also Parker, 1980). Today, similar levels of validity coefficients continue to be reported, such as the .34 coefficient reported in a recent study by Hoffman and Thornton (1997). These validity coefficients are generally statistically significant at the .001 level, and the coefficients reported in the meta-analyses (.37 and .41) compare well with other forms of assessment, including cognitive ability

tests (Coutts, 1990, pp. 110-111; MacKinnon, 1975a, p. 21; Schmitt et al., 1984, pp. 412-416, 420, 429).

In summary, although some explanatory deficiencies exist in assessment theory (Howard, 1974, p. 115; Klimoski & Brickland, 1987, p. 243), which are discussed below, there is considerable evidence that assessment centers have predictive validity, especially for managers, and that it can be generalized across assessment centers (Gaugler et al., 1987, pp. 503-504; Klimoski & Brickner, p. 243; Lowry, 1994; p. 383; Spychalski et al., 1997, p. 85).

Validity Controversies

Initially, for very practical reasons, the focus of assessment center research was not upon the explanation of behavior so much as upon the reliable prediction of behavior in a job-related context. In other words, assessment center research was not concerned with why it worked, just that it did. Recently, however, the focus has been on the "why,"[24] and two thought provoking questions have been raised. The first question challenges the evidence for predictive validity claims, suggesting that the traditional criteria are suspect. The second question challenges the evidence for construct validity claims, suggesting that the traditional explanation is not supported by the evidence.

The Criterion Controversy

One general criticism is that predictive validity studies conducted in operational centers have a design flaw that, although not fatal, casts doubt on the conclusions. The problem is that of criterion contamination, where both the candidates and the employers

[24] For a classic example, see Klimoski & Brickner's (1987) article, "Why Do Assessment Centers Work? The Puzzle of Assessment Center Validity."

are aware of the overall results (MacKinnon, 1975a, p. 18). The question, then, is to what extent do assessment center scores influence the validation criterion (e.g., promotion, or job performance)? Does the assessment center outcome become a self-fulfilling prophecy, either for candidate success or failure (known in the early literature as the "crown prince/princess" effect)? These questions can be answered by comparing the predictive validity coefficients reported in research studies (e.g., MPS) with those reported in studies of operational centers. As reported by Gaugler et al. (1987), validity coefficients reported for operational centers compare well with those of research centers (p. 503), resolving the issue in favor of the assessment center. What is not known, however, is how assessment center participation, usually reserved for those candidates considered to have potential, can change or reinforce self-perceptions, even though candidates may be unaware of the results as is the case in research centers (Klimoski & Brickner, 1987, p. 249, citing Schmitt, Ford & Stultz, 1986).

Another criticism is that which challenges the validation criteria, usually job performance or promotion, and it applies to both operational and research assessment centers. For example, job performance as measured by supervisory evaluations are sometimes criticized because they are often more unreliable than assessment center ratings (Gaugler et al., 1987, p. 494; Norton, 1981, p. 561; see also Schmitt et al., 1984, p. 419). But more importantly, as noted by MacKinnon (1975a), often the validation criteria themselves have not been validated (p. 20). The classic article on this subject is that of Klimoski and Strickland (1977), who suggest that assessment centers may not be identifying the best candidates for the target job; rather, they may be identifying the successful candidates (i.e., those of whom managers will approve). In other words,

Klimoski and Strickland are suggesting that assessment centers may be more "prescient" than valid, turning out clones of what Byham (1980) called the "organization man" (p. 34; see also Hinrichs, 1978, p. 600).

Sometimes known as "subtle criterion contamination," this hypothesis has raised some doubts about assessment center validity. But for the most part, such doubts can be resolved by comparing the predictive power of employee raters with that of consulting raters (e.g., industrial psychologists or employees from other organizations), who have no permanent relationship with the organization in question and have little or no prior knowledge of the organization's culture and expectations. If the "subtle criterion contamination" hypothesis is true, inside employee raters should be able to predict more accurately candidate promotions than consulting raters. However, according to Klimoski and Brickner (1987), research indicates that employees from other organizations obtained validity coefficients comparable to inside staff (p. 248, citing Silzer, 1985). And according to Gaugler et al. (1987), psychologists obtained higher validity coefficients than managers (p. 505). Ironically, the "subtle criterion contamination" hypothesis is itself suspect. Lacking direct evidence (rf. Gaugler et al., 1987, p. 505), it has raised doubts based solely on circumstantial evidence (Klimoski & Brickner, 1987, p. 248) and so should not be considered a serious threat to assessment center validity.

The Construct Controversy

In contrast to the more atheoretical predictive validity, construct validity is theory-laden and "is concerned with the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses concerning the concepts (or constructs) that are being measured" (Carmines & Zeller, 1979, pp. 18, 23). The

construct validity controversy is generally concerned with whether an assessment center measures dimensions (the traditional dimension-based theory) or whether it measures exercises (the new alternative exercise-based theory). As previously discussed in this chapter, the assessment center is based on the theory that dimensions (even complex ones such as leadership, judgment, and problem-solving ability) are reflected in categories of behavior that can be measured across multiple exercises (i.e., situationally through job simulations).[25] However, as summarized by Howard (1997), some research suggests a greater relationship between different dimensions within an exercise than the same dimension across exercises (p. 21). Although there is a relationship between the same dimension measured across exercises (indicating convergent validity), the relationship is generally higher between different dimensions within exercises (indicating a lack of discriminant validity), which suggests that the assessment center may be measuring exercises rather than dimensions.

Credit is generally given to Sackett and Dreher (1982) for first challenging the traditional dimension-based theory (Kleinman & Koller, 1997, p. 65; Reilly, Henry & Smither, 1990, p. 82). Sackett and Dreher argue that it is an unsupported assumption, providing empirical evidence in support of exercise-based theory (p. 409; see also Bycio, Alvares & Hahn, 1987; Donahue, Truxillo, Cornwell & Gerrity, 1997, p. 86; and Lowry, 1997, p. 54).[26] It is not suggested that assessment centers do not work (i.e., in terms of criterion-related validity), but that they do not work as explained by traditional theory

[25] Byham (1977a) uses the analogy of aerial mapping, where photographers (raters) using overlapping pictures (exercises) in aerial mapping in order to construct the larger picture (dimension).
[26] These authors provide additional evidence in support of Sackett and Dreher (1982), and also review research conducted since 1982 that supports the exercise-based theory.

(Klimoski & Brickner, 1987, p. 246). Rather than dimensions, exercise-based theory claims that the proper focus of assessment centers should be the exercises, which are defined as job samples or job-related tasks, and which are considered to be the real basis for making judgments. Therefore, if the exercises are in fact unidimensional tasks, then any attempt at identifying underlying dimensions is simply misguided (Lowry, 1997, p. 55; see also Klimoski & Brickner, pp. 251, 254; and Roberton, Gratton & Sharpley, 1987). In an apparent attempt to accommodate both theories, the 1989 Guidelines for assessment centers permit exercises to be used as an alternative to dimensions (p.2).

From the perspective suggested in the Guidelines (1989), Howard (1997) argues that although using tasks instead of dimensions may be appropriate for some occupations, it creates a new set of problems for others (p. 27). Different jobs require different abilities that may be described on a continuum—from manual labor jobs, where competency may be measured by observing the performance of physical tasks, to professional jobs, where competency can only be measured through observing complex behaviors. For example, the policing profession requires a complex set of skills, knowledge, and abilities, such as using discretion when enforcing the law. If such a competency were arbitrarily reduced to a physical task, most relevant behavior would be lost (Jones, 1997, p. 174). Regardless of complexity, however, according to traditional theory, dimensions may be defined as "a description under which behavior can be reliably classified" (Byham, 1980c, p. 29; see also Norton, 1981, p. 564).

In addressing evidence of low discriminant validity in the traditional assessment center, Howard (1997) argues that it may be partly attributable to the recent unconventional practice where raters assess dimensions within an exercise, which

artificially produces with-in exercise bias (p. 26; see also Thornton, Tziner, Dahan, Clevenger & Meir, 1997, p. 111-112, 126). This practice is contrary to the original design of the assessment center as described in MPS (which Sackett and Dreher (1982) acknowledge), where raters record their behavioral observations in each exercise, but suspend assessing dimensions until all observations are complete (Adams & Thornton, 1988, p. 4). As noted by Thornton et al. (1997), those analyses that find against the dimension-based theory "have not made the proper distinction between these methods" (p. 110).[27]

It is this within-exercise scoring practice that yields the empirical evidence upon which critics base their conclusion that the traditional dimension-based theory should be abandoned. The problem, though, is that evidence of low discriminant validity depends upon statistical analyses (i.e., primarily the multitrait-multimethod matrix approach,[28] and to a lesser extent conventional factor analysis) that require within-exercise assessments, thus confounding the results (and also creating an interesting methodological dilemma). Citing Joyce, Thayer, and Pond (1994), Howard (1997) suggests that "perhaps it is time to conclude that traditional construct validation techniques are inappropriate for assessment centers" (pp. 26-27; see also Jones, 1997, p. 175; Kleinman & Koller, 1997, pp. 66-67, 80-81; and Reilly et al., 1990, p. 83).

---

[27] For operational assessment centers that primarily identify suitable candidates for hiring or promotion, overall ratings are generally the ratings of interest, except in the case of developmental assessment centers, where participants are expecting feedback on individual dimensions (see Kudisch, Ladd & Dobbins, 1997).

[28] The multitrait-multimethod matrix approach (MTMM) was proposed by Campbell and Fiske (1959) for assessing construct validity, and is very widely used. Kleinman and Koller (1997) provide a brief, conceptual overview of the method, along with its limitations and alternative statistical approaches, such as confirmatory factor analysis (pp. 66-67; see also Adams & Thornton, 1988, pp. 10-12).

For example, according to Thornton et al. (1997) factor analysis has produced factors, or domains, for dimensions that are inter-related (p. 112; 116-118, 125; see also Kudisch, Ladd & Dobbins, 1997, p. 139). Subsequently, construct validity studies have correlated these domains, such as a cluster of interpersonal dimensions, with external measures of related constructs (Howard, 1997, p. 24; Thornton et al., p. 113, 125). Notably, in the original assessment centers, exercises were selected to represent major domains and not necessarily every dimension (Howard, p. 24). Consequently, an exercise can be expected to include similar dimensions, rather than a random collection of dimensions, within the same domain and so correlate highly within that exercise.

In addition to being an artifact of the analysis, research has indicated that poor construct validity of dimension ratings may also be attributable to natural human limitations. Research has shown that numerous and complex dimensions may result in unrealistic cognitive demands being placed upon raters, who simply cannot accurately identify and categorize all observed behaviors into the appropriate dimensions (Gaugler & Thornton, 1989, p. 611; Reilly et al., 1990, p. 72-73; see also Adams & Thornton, 1988; and Bycio et al., 1987). Additionally, the dimensions may be ambiguous and overlap, and the exercises may be inadequate to extract the dimensions of interest, at least to the extent necessary for raters to recognize and categorize them. By reducing dimensions from the traditional eight or more to three and using behavioral checklists, Reilly et al. found that convergent validity was substantially increased (.43, compared to the average of .24 found in critical studies), and the discriminant validity coefficient improved (.41), falling below that of the convergent validity coefficient.[29]

[29] Donahue et al. (1997) found in for exercise-based theory, but they used "untranslated" behavioral checklists (responses are not coded by subject experts), while Reilly et al. used "translated" checklists.

In summary, although Klimoski and Brickner (1987) challenged the dimension-based theory, they warned that evidence for the exercise-based theory was tentative (p. 255); and ten years later, Lowry (1997) reported that exercise-based research was still limited (p. 61). To date, as noted by Howard (1997), studies indicating problems with construct validity, particularly discriminant validity, are not sufficiently convincing to conclude that the traditional dimension-based theory should be abandoned or replaced by the alternative exercised-based theory (p. 22). Byham bluntly states that exercise domains "just don't work out in practice" (quoted in Mayes, 1997, p. 12). This is not to suggest that the debate on construct validity is closed—theoretically, the issue is somewhat troublesome. However, it is important to put the matter into perspective. Construct validity is a complex concept under the best of circumstances and will not likely be resolved anytime soon in the context of the assessment center.[30] And for operational purposes, the debate is academic because the traditional assessment center method for over forty years has produced acceptable levels of criterion-related validity along with comparatively low levels of discrimination.

Police Use of the Assessment Center

This literature review will now conclude by focussing on police use of the assessment center method, which will more clearly specify the context in which this study occurs.

---

[30] According to McGinnis (1987), construct validity is an idea that was only introduced into psychometric literature in 1955 by Cronbach and Meehl, who defined it as "some postulated attribute of people assumed to be reflected in test performance" (cited at p. 98). A great deal of ink has been spilled in the literature in an attempt to clarify this "simple but slippery idea" (Moore, 1985, p. 114), which Tenopyr (1977) describes as "construct-content confusion."

History

In the United States, minimum criteria for the formal selection of police officers first appeared in 1870, when President Ulysses S. Grant established standards for government positions (Gavin & Hamilton, 1975, p. 166, citing Holmes, 1942, p. 515). For the next 100 years police selection was generally traditional; but similar to the private sector, in the 1960s and 70s there was an emerging trend toward the use of multiple selection procedures and situational testing (Gavin & Hamilton, pp. 167-168, 171). During this time of experimentation, the MPS assessment center, which combined both multiple selection procedures and situational testing, gained momentum and became the dominant model for behaviorally focussed personnel selection procedures, whether in the public sector (e.g., police) or the private sector (e.g., AT & T).

In 1970, the International Association of Chiefs of Police (IACP), one of the leading proponents of the assessment center method for police selection, initiated efforts to establish an assessment center in the Mississippi Bureau of Narcotics (Quarles, 1982). It was the position of the IACP that paper and pencil tests for recruit selection were inadequate, and that the assessment center offered more potential for identifying the best candidates for policing (McGhee & Deen, 1979). Since then, the IACP has continued to support the use of the assessment center method by sponsoring assessment centers and by conducting training conferences and publishing articles on the assessment center method (Kohlhepp, 1992).

According to Tielsch and Whisenand (1977), England was the first country to use the assessment center method for police selection (p. 25). In the United States, the first reported experimental application of the assessment center method by municipal

governments was in police and fire departments in the early 1970s (Adair & Moon, 1977; Byham, 1977a, p. 41; Maher, 1984, p. 21; Stinchcomb, 1985). The historical record for police use is somewhat uncertain. For example, Quarles (1982) states that the IACP in 1971 established the first pilot assessment center (noted above), Byham (1977a) states that New York City Police was the first to apply the method (although it was restricted to identifying upper management), and Gavin and Hamilton (1975) state that the police in Fort Collins (Colorado) and Colorado State University were first. Regardless of the historical record, evidence for assessment center validity resulted in other cities adopting the method to select police, including St. Louis, Kansas, Rochester, Richmond, and the Federal Bureau of Investigation (FBI) (Brown, 1978, p. 61).

The assessment center method has continued to experience rapid growth in law enforcement agencies in the United States. For example, in a paper presented to the International Congress on the Assessment Center Method, O'Leary (1997) reported that approximately 32% of police departments in the United States use the assessment center method, which is similar to that reported by Lowry (1994; 1996; see also 1997) and Fitzgerald and Quaintance (1982). Despite widespread use, however, there is little published research (excluding the brief descriptive articles that appear regularly in law enforcement magazines) on police use of the assessment center method. For example, Pynes and Bernardin in 1989 claimed to have conducted the first empirical predictive validity study on police recruitment, and Ross in 1980 claimed to have conducted one of the first predictive validity studies on police managerial job performance (other validity studies include those by Feltham in 1988, O'Hara & Love in 1987, Gavin & Hamilton in 1975, and Mills in 1976).

In Canada, consistent use of the traditional assessment center method by police for hiring purposes is limited to the province of British Columbia, where since 1978 most of the 12 municipal police departments have used the assessment center operated by the provincial Police Academy. As might be expected, most published literature in Canada on police use of the assessment center method is that published on the Police Academy, and it is mostly descriptive in content (e.g., Gale, 1983; McGinnis & Carpenter, 1980; Taylor, 1983; Turner, 1978; and Turner & Higgins, 1977).

Reliability

With respect to reliability, the only published research relevant to policing is a legal case, which is believed to be the first legal challenge in the United States of a police assessment center on the grounds of fairness or due process (Mendenhall, 1989; 1992; Hurley, 1987, pp. 38-39). This case (Berry v. Omaha, 1975), which Fitzgerald and Quaintance (1982, p. 14) refer to as the "classic case" addressing procedural issues in a police assessment center, involved a group of unsuccessful promotional candidates who claimed that the assessment method was unfair and so took the city of Omaha to court. Byham, who had audited the assessment center results, testified for the city. He argued that because the assessment center demonstrated reliability and adherence to the standards established by the International Congress on the Assessment Center Method, it was a fair and valid process. Specifically, Byham had used correlational analysis (Spearman's "rho") and reported a coefficient of .84, which was described as "very high" (Mendenhall, 1992, p. 62). The court accepted Byham's evidence and found that the assessment center was not arbitrary and capricious, concluding that such testing procedures assess candidates fairly (Mendenhall, p. 63; Hurley, p. 38).

Discrimination

There is little published research on discrimination in procedures used for police recruitment and selection,[31] and less yet on discrimination in assessment centers used for police selection.[32] In one study, Magaldi, Mendoza, Stafford and Frank (1984) analyzed data from a promotional assessment center for the Metro-Dade Police Department. Their basic research question was whether there was "evidence of discrimination or adverse impact" on the grounds of sex or race. After an analysis of assessment center scores, they found no statistically significant differences between candidates with respect to race or sex. Similarly, they found no adverse impact based on the four-fifths rule, concluding that officers who participated in the assessment center received a "fair, unbiased evaluation of their skill" (pp. 15-16).

In another American study, Pynes and Bernardin (1992) analyzed four years of data (1982-1986) from a recruit assessment center. In this study, using the four-fifths rule (Equal Employment Opportunity Commission, 1978), the authors compared written cognitive ability examinations with the assessment center to determine which method resulted in less racial discrimination. If only cognitive ability examinations were used in the selection process, black and Hispanic selection rates, compared to the white selection rate, were .47 and .77 respectively. However, if only the assessment center were used, black and Hispanic selection rates were .63 and .86 respectively. In spite of the fact that the assessment center method was more expensive than traditional written examinations,

[31] This study did not focus on the use of assessment center for promotion.
[32] For one of the first articles on discrimination in police selection, see Balzer's (1976) article entitled, "A View of the Quota System in the San Francisco Police Department."

the authors recommended the assessment center as a personnel selection process because discriminatory effect was reduced (cp. Hoffman & Thornton, 1997).

In Canada, published literature on selection fairness is "virtually non-existent" (Singer, 1993, p. 30), which has been noted in a number of Canadian studies on policing (e.g., Linden & Minch, 1994; Nelson, 1992; Policing in British Columbia Commission of Inquiry, 1994; and Walker, 1993). Moreover, there are no specific studies in Canada on reliability in assessment centers, except for two in-house studies by the Police Academy (Chamberlain, 1980; Collins, 1985). Quite accurately, Nelson (1992) described the lack of Canadian research on discrimination in police selection as glaring (p. 193).[33]

Validity

Using job performance as the validation criterion, Ross (1980) reported a validity coefficient of .47 in a study of five police agencies in the United States. Similarly, Feltham (1988) reported a validity coefficient of .33 in a British study. In the same study, however, using promotion as the validation criterion, Feltham reported negligible validity. In a police recruiting assessment center, Pynes and Bernardin (1989) reported low validity coefficients of .14 and .20 on the criteria of police academy scores and on-the-job performance ratings respectively (see also Pynes, 1988). The findings of the latter two studies are contrary to those found in the literature on assessment centers in general, and the authors suggest that confounding variables might explain the differences.[34]

---

[33] As a result, contextual information is missing, which in turn is a limiting factor when the results of the discrimination analysis are reported in Chapter 6.

[34] See also Hirsh, Northrop, and Schmidt (1986), who conducted a meta-analysis of validation studies with respect to the use of cognitive ability tests by police. A low validity coefficient of .13 for job performance (compared to .36 for training success) was explained as possibly the result of the police profession's dependence upon personality-type variables, such as interpersonal skills, for success (cited in Pynes and Bernardin, 1992, p. 42).

Both Feltham and Pynes and Bernardin noted that range restriction caused by high ability levels may have reduced the validity coefficient, and Pynes and Bernardin also noted that the complexity of police work might present more difficulties for job analysis. Additionally, Pynes and Bernardin noted that the assessment center in their study used a three-point rating scale as opposed to the more traditional five-point rating scale, which has been shown to produce better reliability (Landy & Farr, 1980, cited by Feltham, 1988, p. 134). Finally, Feltham cautioned against comparing his study to those in the United States that are mostly modeled after MPS. British assessment centers, often called "extended interviews" (EI's), are generally not modeled on MPS and so may have significant differences, such as the absence of role-playing.

Coutts (1990), as part of a larger study on police hiring and promotion in Canada, briefly addressed the assessment center method. He concluded that "one of the most significant advancements in the police personnel selection field in ... [the United States] has been the increased application of the assessment center method" (p. 110; see also Hurley, 1987, p. 24). Coutts, in his review of validity literature on assessment centers, found that as a recruit selection tool the assessment center is comparable to the best alternative techniques (e.g., cognitive ability tests) and superior for promotional selection (p. 111). Although evidence on the use of the assessment center method within the policing profession is inconclusive due to a lack of research (McGinnis, 1987, p. 108), as previously discussed there is considerable evidence for the predictive validity of assessment centers in general (e.g., see Gaugler et al., 1987, pp. 503-504).

Summary

The purpose of this chapter was to review the literature on the assessment center method. In each of the sections a particular emphasis was placed on theory and history, which provides the framework in which to understand the operation of an assessment center. Moving from a general orientation, the review narrowed its focus to assessment centers dedicated to the selection of police, either for job entry or for promotion. As a result, this chapter has provided the necessary background and context for the following discussions: interrater reliability in Chapter 3, discrimination in Chapter 4, and Police Academy assessment center operations in Chapter 5.

It is important to note, however, that although the Police Academy assessment center is the object of this study, as discussed in Chapter 1, the overall concern is that of substantive equality as defined by law and the goal is to provide a model in which systemic discrimination in an assessment center can be assessed and prevented. Therefore, issues relating to personnel selection, assessment center methodology, and measurement must be interpreted within a legal framework, which defines the concept of discrimination as presented in this study. Here, the relevance of facts (e.g., the results of statistical tests) is often a "question of law," which will be clarified in the discussion on discrimination and as this study proceeds.

CHAPTER III: INTERRATER RELIABILITY

> Reliability is a major concern ... when a psychological test or questionnaire is used to measure some attribute or behavior. If we are to understand the functioning of a test, we must understand its reliability, i.e., the extent to which it consistently discriminates individuals at one time or over the course of time [emphasis in original]. (Rosenthal & Rosnow, 1991, p. 47)

Introduction

In order for a test (or any measuring procedure) to be useful, it must be reliable and valid. Reliability indicates the degree to which a test yields consistent results and high reliability is essential for accurate discrimination between subjects on a dimension of interest (characteristic or variable) by means of a standardized scale. This valid form of discrimination is distinguished from "observer bias" (Borg & Gall, 1983), which is a reliability error that will invalidate a test and may be unjustified in the employment context. Validity indicates the degree to which a test is theoretically coherent and high validity is essential for confidence that the test measurements reflect the dimension that they are intended to measure. In other words, reliability focuses on a particular property of a test, while validity focuses on its theoretical claims (Carmines & Zeller, 1979, pp. 11-12; Kachigan, 1991, pp. 139-140).

Reliability is a pre-requisite for validity—logically, a test that produces inconsistent results (measurements) cannot provide valid information about the dimension of interest (Gronlund, 1985, p. 87; Kachigan, 1991, p. 141; Rosenthal & Rosnow, 1991, pp. 47, 50). In such a case, an assessment of test validity is impossible because the test score or predictor cannot be correlated with the validation criterion (Kaplan & Saccuzzo, 1982, p. 91). Specifically, "a test cannot correlate better with an

external variable than it does with its own true score" (Cronbach, 1984, p. 176).[1] It is important to note, then, that validity does not necessarily follow reliability, but reliability provides the consistency that makes validity possible.

Shrout and Fleiss (1979) point out that judgments made by humans are especially plagued by problems of reliability (p. 420). This problem is exacerbated by confusion in the literature. For example, in an article on test measurement and reliability, Li, Rosenthal and Rubin (1996) state that "psychologists are becoming so specialized and so differentiated that at times we find it difficult to understand one another's literature" (pp. 98, 100). But, as noted by Rosenthal and Rosnow (1991), "A research idea should not be any more complicated than necessary" (p. 32). To make their point, Rosenthal and Rosnow use the metaphor of "Occam's razor." Occam was a fourteenth-century philosopher who believed that "what can be explained on fewer principles is explained needlessly by more" (Rosenthal & Rosnow, pp. 32, 89). In other words, unnecessary and unwieldy details should be cut away. The rationale of this metaphor, then, is not to justify an overly simplistic analysis but rather to avoid an unnecessarily complicated analysis when a simpler one would do just as well (Rosenthal & Rosnow, pp. 32, 391).

Considering that distinguished academics such as those noted above find it difficult to understand some of the specialized literature, it seems reasonable to conclude that many field practitioners probably find the subject quite overwhelming. For example, as noted in Chapter 2, in a recent survey of assessment center practices in the United States, 21.4% of respondents reported that they did not evaluate reliability (Spychalski,

[1] Because validity is usually measured by correlating two variables (the predictor and the criterion) with each other, validity is in part a function of the reliabilities of the two variables in question. This relationship will be discussed in more detail in the section on "Correction for Attenuation."

Quinones, Gaugler & Pohley, 1997, p. 80). Such a situation is quite unacceptable because important decisions affecting people are based on rater assessments (Kaplan & Saccuzzo, 1982, p. 90). As noted by Cronbach (1984), when making decisions about people, a measurement process with unacceptable levels of error, especially systematic error, may be found to be unjust, unlawful, and unprofessional (p. 159; see also Cronbach, Yalow & Schaefer, 1980, p. 693). Consequently, the importance of reliability to the measurement of people cannot be overstated.

With this in mind, the goal of this chapter is to apply Occam's razor to the concept of interrater reliability measurement in assessment centers. This is especially important for two reasons. First, because a generalized reliability coefficient cannot be calculated for the assessment center method, it is essential that administrators be able to estimate interrater reliability for their particular assessment centers (Tinsley & Weiss, 1975, p. 373).[2] Second, because discrimination can be effectively measured by a correlational analysis (see Chapters 5 and 6), knowledge of reliability can be useful for estimating the potential for discrimination by correcting correlations for unreliability. The objectives of this chapter are to (1) clarify important issues in measurement theory, (2) review the most common methods for assessing interrater reliability, and (3) recommend a practical and uncomplicated yet suitable parametric method for estimating interrater reliability. As a result, assessment center administrators will hopefully have a better understanding of the concept of reliability and recognize its importance in both behavioral assessment and discrimination assessment.

[2] In contrast to well-developed written tests that can be strictly standardized, the assessment center method depends upon critical factors (e.g., the selection, training and aptitude of raters and the construction of simulations) that often differ substantially from assessment center to assessment center.

The major limitation of this study is that it focuses on the most common parametric indices of reliability that measure consistency. First, being limited to parametric indices, only brief attention is paid to non-parametric indices, such as Spearman's rho, Kendall's *W*, and Cohen's *kappa*, which are generally used to estimate the reliability of measurement of nominal and ordinal data. However, assessment centers yield data that can be considered to be at an interval level—behavioral rating scale intervals are assumed to be approximately equal and the underlying dimensions are theoretically continuous in nature (rf. Kachigan, 1991, pp. 13-14; 16). Consequently, focussing on parametric tests of reliability is arguably suitable for the purposes of this study. Second, being limited to the most common parametric indices, other parametric indices such as multivariate procedures (rf. Martin & Bateson, 1993, p. 124; Rosenthal & Rosnow, 1991, Chapter 24; Tinsley & Weiss, 1975, p. 365, citing Overall, 1965) were not reviewed. Moreover, these procedures are complex and are not widely used in operational assessment centers. Third, being limited to parametric indices that measure consistency, measures of agreement, such as and Tinsley & Weiss' (1975) T-index, were not reviewed. These indices, however, are not generally used for numeric data unless absolute agreement is an issue. Finally, the study of reliability is much broader than what can be presented in this chapter, which is necessarily limited by issues of practicality.

Measurement Theory

According to the Bible, before sentencing Jesus to death by crucifixion, Pontius Pilate asked him, "What is truth?"[3] Philosophers have been intrigued by this simple yet

[3] John 18:38.

profound question throughout history, regardless of their fields of interest. For example, mathematicians write about the "true" score when debating classical measurement theory. But just as Jesus left Pilate wondering about the meaning of truth, mathematicians leave us wondering about the meaning of the true score. Predictably, there are many unsettled issues in the field of reliability theory.

There is, however, one point regarding reliability upon which all mathematicians agree: all measurements are subject to error. For example, with respect to measurements made on people, Cronbach (1984) said that "no single observation fully represents the person" because no procedure is completely trustworthy (p. 158). Many reasons may explain variation in measurement, such as changes in the person when stability over time is the consideration, unaccountable chance variation, and systematic variation. This raises the issue of a single measurement, which, for practical reasons, is how many judgments are made about people on important dimensions such as intelligence and aptitude. Here, there are no guarantees that the single score obtained by a particular test is a true reflection of the characteristic or behavior of interest. For this reason, objective information about the reliability of measurement scales or tests is essential in order to make informed decisions about the usefulness of a test and whether it should be used as a basis for making judgments about people.

Reliability Theory: A Brief History

The idea of chance or random sampling error, based on probability theory, was introduced in the 17th century, and the product-moment correlation was introduced in 1896 by Karl Pearson. Together, these two concepts form the basis of measurement in reliability theory. The seminal works in reliability theory were published in 1904. One

was an article by Charles Spearman, "The Proof and Measurement of Association between Two Things," and the other a text by Edward Thorndike, Introduction to the Theory of Mental and Social Measurements, both of which form the foundation of contemporary reliability theory (Kaplan & Saccuzzo, 1982, p. 87; Traub, 1994, p. 1).

In 1910 Spearman and Brown simultaneously, yet independently, developed one of the most basic theories of reliability (Carmines & Zeller, 1979, p. 41; Rosenthal & Rubin, 1982, p. 98; Rosenthal & Rosnow, 1991, p. 48). That is, if the reliability of test items (known as internal consistency reliability or reliability of components) is a function of the arithmetic average of the intercorrelations between all test items, and if all test items are positively correlated, then reliability will increase as test items are increased (Li et al., 1996, p. 98; Rosenthal & Rosnow, 1991, p. 48; see also Tinsley & Weiss, 1975, p. 365, citing Ebel, 1951). The equation developed by Spearman and Brown (now known as the Spearman-Brown correction formula) to estimate reliability after adding test items is as follows:

$$R = \frac{n\bar{r}}{1 + (n-1)\bar{r}},$$

where R is the corrected reliability coefficient, n is the correction factor (i.e., the factor by which the test is lengthened or shortened), and $\bar{r}$ is the mean correlation among all items.

Classical Reliability Theory[4]

Classical reliability theory is known by a number of different names, such as classical test theory and true score theory. According to Cronbach (1984), this theory

[4] This discussion of classical reliability theory is only intended to provide an elementary introduction to the subject. For a comprehensive analysis, see Nunnally's (1978) text, Psychometric theory, or more recently, Traub's (1994) text, Reliability for the social sciences.

defines "error" as unwanted variation (p. 159). Measurement of any target (e.g., rating a person's behavior) results in an observed score and is subject to error. The problem is in determining the amount that the observed score differs from what is theoretically known as the "true score." Because the true score cannot be known, it is often redefined as the "expected value of a random variable" (Traub, 1994, p. 19).

Measurement error, as explained by classical test theory, is the difference between the observed score and the true score, and a major assumption is that measurement error is random (Carmines & Zeller, 1979, p. 30; Kaplan & Saccuzzo, 1982, p. 88; Traub, 1994, pp. 24-25). Such errors can either be positive or negative, and classical test theory assumes that over repeated measurements, errors due to random chance will negate each other or average to zero (Cronbach, 1984, p. 159; Traub, p. 25). Notably, the amount of random error is inversely related to the degree of reliability, which will become clear as this discussion proceeds. Nonrandom errors do not cancel out over repeated measurements and as a result threaten "the very heart of validity" by introducing unknown systematic error or bias that prevents the test from measuring what is intended (Carmines & Zeller, pp. 14-15). Systematic error may be caused by any number of factors, from the measurer to the measurement instrument to that being measured. Notably, systematic error does not necessarily adversely affect the degree of reliability, and in fact may enhance it.

Given the assumption of random error, if one subject were measured one time on a dimension of interest, then the basic formula for measurement reliability would be as follows:

$$X = t + e,$$

where X is the observed score, t is the true score, and e is the random error. However, as previously noted, the true score is hypothetical and cannot be known, but it can be estimated if the subject were measured numerous times. As random errors are assumed to cancel each other out and average to zero, it follows that the expected value or average of the observed score (within a certain confidence interval) for the subject should equal his or her theoretical true score (Carmines & Zeller, 1979, p. 30; Traub, 1994, p. 25):

$$\varepsilon(X) = t,$$

where $\varepsilon(X)$ is the expected long term average of the subject's observed score and t is the hypothetical true score.

Even though the same test is used on one subject, due to random error numerous measurements will produce many different observed scores that result in a sample probability distribution (vis-a-vis sampling probability distribution). From this distribution a mean and standard deviation can be calculated. In classical test theory, the mean is assumed to approximate the true measurement score on the subject, as explained above. Associated with the observed scores are errors, which also result in a sample probability distribution. The standard deviation of this distribution is commonly known as "the person-specific standard error of measurement" (PSEM), to distinguish it from the "standard error of measurement" for a population (SEM),[5] and the square of the PSEM is commonly known as "error variance." Of particular significance is that the shape of the

[5] The PSEM and SEM should not be confused with the standard error (SE), which is the sampling distribution of the mean.

observed scores probability distribution is the same as that of the errors distribution. As a result, the standard deviation of the observed scores is identical to the standard deviation of the error scores (i.e., PSEM) (Cronbach, 1984, p. 159; Traub, 1994, pp. 25-28; see also Kachigan, 1991, pp. 95-96).[6] Consequently, the size of the PSEM is a measure of the accuracy of the estimated true score for a person.

The problem is that this theory only addresses the error distribution for one person (i.e., it is "person-specific"), and it may not be practicable to repeatedly measure the same person to estimate his or her individual PSEM. Instead, many persons are measured once and the measurement distribution is assumed to approximate the distribution of repeated measurements of a single person[7] (Kaplan & Saccuzzo, 1982, p. 89; Traub, 1994, p. 39). Specifically, once the reliability coefficient and observed score variance are known for a population, a "standard error of measurement" (SEM) can be estimated, which is assumed to approximate the error distribution of repeated measurements of a single person. As a result, the size of the SEM is a measure of the accuracy of the estimated true score for an individual (rf. footnote 9).

Since reliability conventionally refers to the consistency of repeated measurements across many targets rather than within a single target, the equation $X = t + e$ noted above must be rewritten so that it does not apply to a single occurrence of an observed score, true score and error. Rather, it must apply to the variance of those properties that occur when a single measurement is made of many persons on a

---

[6] This is logical because adding a constant value (e.g., the mean of the observed scores) to each error affects neither the standard deviation or distribution shape. Similarly, the shape and standard error (SE) of the sampling distribution of a mean $\bar{x}$ is identical to the shape and standard deviation (s) of the related error distribution.

[7] More precisely, the error variance here is the expected value of the variance of the error for a single person (Traub, 1994, pp. 20, 25-26, 34).

dimension of interest (Carmines & Zeller, 1979, p. 31). Given the assumption of random errors, the following fundamental equation of classical reliability theory should also be true (Kaplan & Saccuzzo, 1982, pp. 87-88; Traub, 1994, pp. 19, 24, 32):

observed score variance = true score variance + error variance
(true difference between persons) (assumed random chance)

Therefore:

true score variance = observed score variance – error variance,

and

error variance = observed score variance – true score variance.

Logically, then, the ratio of true score variance to observed score variance may be called the reliability of the measure, which indicates its consistency. In other words, this ratio indicates the percentage relationship between true score variance and observed score variance (Cronbach, 1984, p. 160; Kaplan & Saccuzzo, 1982; p. 90; Kozlowski & Hattrup, 1992, p. 164; Traub, 1994, p. 38):

$$r_{xx1} = \frac{\text{true score variance}}{\text{observed score variance}}.$$

The reliability ratio can also be expressed in terms of the error variance (Carmines & Zeller, 1979, p. 31):

$$r_{xx1} = 1 - \frac{\text{error variance}}{\text{observed score variance}}.$$

This reliability ratio is the basis for the application of reliability formulae (e.g., intraclass formulae, described in detail later). As the true score cannot be known, it must be estimated from the observed score. If error variance is due to random chance then it can be estimated, and true score variance may be written as "observed score variance – error variance." Consequently, the reliability coefficient may be expressed as follows:

$$r_{xx_1} = \frac{\text{observed score variance - error variance}}{\text{observed score variance}}.$$

As observed score variance (which logically exceeds true score variance as it consists of true score variance plus error variance) approaches true score variance, the reliability coefficient (which is theoretically bounded by 0.0 and 1.0 in this ratio)[8] approaches a coefficient of 1.0. Therefore, when observed score variance equals true score variance, no error exists and the reliability coefficient (or correlation) is a perfect 1.0 (Traub, 1994, p. 38).

The symbol $r_{xx'}$ is traditionally used because a reliability coefficient is really a function of the correlation of one measure x with another measure x_1 (Cronbach, 1984, p. 160), as discussed above in the section on history. Specifically, the reliability of a measure may be estimated by correlating parallel measurements of that measure (Carmines & Zeller, 1979, p. 32). Measurements are strictly parallel if they have the same true scores and equal (random error) variances, and "reliability can be expressed in terms of the variances of these properties" (Carmines & Zeller, p. 32). Because any

[8] This is important to note, especially when interpreting the results of a reliability analysis (see Chapter 6).

differences between the two scores are assumed to be the result of random error, and the standard deviations are assumed to be equal, it can be shown that the correlation between parallel measures is equal to the true score variance divided by the observed score variance (Carmines & Zeller, p. 33):

$$r_{xx_1} = \frac{\text{true score variance}}{\text{observed score variance}}.$$

It will be noted that this ratio is the same as the reliability formula provided above, the significance of which is that the hypothetical true score variance can be estimated from the observed score and correlation between parallel measures (i.e., the reliability coefficient) (Carmines & Zeller, 1979, pp. 32-33):

$$r_{xx_1} \times \text{observed score variance} = \text{true score variance}.$$

Consequently, the true score variance can be subtracted from the observed score variance to estimate the error variance,[9] specifically:

$$(1 - r_{xx_1}) \times \text{observed score variance} = \text{error variance}$$

In conclusion, although classical measurement theory defines reliability as the ratio of true score variance to observed score variance, operationally a reliability

[9] Notably, SEM equals the square root of the error variance and so can be estimated from this formula; i.e., calculating the square root of the left side of the equation $(SEM = s\sqrt{1-r})$ (rf. Cronbach, 1984, p. 160; Gronlund, 1985, pp. 96, 99; and Traub, 1994, p. 41). As previously noted, SEM provides a confidence band for interpreting the accuracy of an observed score on a single person. For example, based on probability theory (rf. Moore, 1985, Ch. 8, "Probability: The Study of Randomness"), given a person's observed score x_o, one can be confident that 68% of the time the observed score x_o will be one SEM from the person's theoretical true score x_t (i.e., 68% of the time, $x_t = x_o \pm 1SEM$) (rf. Traub, p. 42).

coefficient will be some type of estimate of the self-correlation of a test (Guilford & Fruchter, 1978, p. 411), which will be the focus of the remainder of this chapter.

Reliability, Agreement, and Consistency

Rating scales "generally require the rater to make a judgment about some characteristic of an object by assigning it to some point on a scale defined in terms of that characteristic" (Tinsley & Weiss, 1975, p. 358). This definition defines precisely the activity of raters at an assessment center—they are required to make a judgment on a subject's behavior by assigning it to a point on a rating scale that has been defined in terms of behavior performance. When raters make such judgments on the behavior of various subjects, their judgments should be reliable; but does that mean consistency, agreement, consensus, or concordance? The interpretations of these terms have caused a great deal of debate in reliability theory and the issues are far from settled. So, before proceeding, it is necessary to review the debate and hopefully clarify the meaning of these terms, at least for the purposes of this study.

The debate

A good example of the debate is provided in the attempt by James, Demaree and Wolf (1984) to propose a new formula (r_{wg}) to estimate within-group interrater reliability on one variable on a single target. Schmidt and Hunter (1989) responded very critically to the proposal by James et al., arguing that such an index of interrater agreement was "illogical, uninterpretable, and meaningless" (see also Kozlowski & Hattrup, 1992, p. 161; and James et al., 1993, p. 306). According to Kozlowski and Hattrup, who came to the defense of James et al. (1984), the dissension was largely rooted in confusion over the meaning of the terms "reliability," "agreement," and "consistency" (pp. 161-162).

This confusion has a long history. For example, in the last 25 years some researchers began to interpret the term reliability as meaning consistency, which they defined as the "proportional consistency of variance among raters ... and is correlational in nature" (Kozlowski & Hattrup, 1992, pp. 162-163). In other words, reliability was defined in terms of consistency in the relative judgments of raters, and is usually reported by way of parametric indices, such as product-moment correlation or analysis of variance (Tinsley & Weiss, 1975, p. 359). Agreement was defined as the extent to which raters make essentially the same absolute ratings, thus assuming that raters are theoretically interchangeable (Kozlowski & Hattrup, p. 163; Tinsley & Weiss, p. 359). As a result, some researchers have argued that agreement is conceptually distinct from reliability (notwithstanding that agreement continues to be classified under the general heading of reliability in all major classification indices).

Continuing with the debate, in 1993 James et al. published an article supporting their original position. They stated that their original article defined agreement as a "special form of interrater reliability," similar to the concept of interchangeability described by Shrout and Fleiss (1979) (p. 306). Under the broader heading of reliability, then, interrater consistency is a correlational index (consistent with classical test theory) while interrater agreement is an interchangeability index (James, et al., 1993, p. 306). Moreover, James et al. argue that their index (r_{wg}) is in fact derived from "classic measurement theory as an interchangeability (agreement) index of interrater reliability," although they recognize this may strain classical reliability theory (p. 306).

Nevertheless, James et al. (1993) conceded to their critics and recast their index as purely one of agreement, without relying on classical reliability theory because they

believed that the debate defeated the purpose of a necessary statistic (p. 307). This concession, however, is not new in the field, as alternatives to classical reliability theory include generalizability theory (Cronbach, Gleser, Nanada & Rajaratnam, 1972) and item response theory (Lord, 1980, cited by Traub, 1994, p. 3). These alternatives may concentrate on a discussion of measurement characteristics that may apply to only a single statistic that must stand or fall on its own merits (such the statistic r_{wg} proposed by James et al.).

Such concessions, though, have not settled the debate, which ranges from the position that "some forms of agreement" are properly defined in the context of interrater reliability to the position that they are conceptually distinct and so unrelated to classical reliability test theory (James et al., 1993, p. 307). For example, in the first position, reliability is defined generally as interrater consistency, by which raters rank subjects proportionately in terms of level, and may include the concept of interchangeability (a form of agreement) between raters. In the second position, agreement is defined exclusively as interrater consensus, by which raters assign the same (or nearly the same) absolute values to the subjects (Fleenor, Fleenor & Grossnickle, 1996, p. 368).

The idea of reliability and agreement being conceptually distinct is supported by the fact that consistency may be high while agreement may be low, and vice-versa (Fleenor et al., 1996, p. 368; Tinsley & Weiss, 1975, pp. 359-360). This phenomenon, however, may be explained in part by methodological limitations. For example, interrater consistency is usually estimated by means of the product-moment correlation or analysis of variance, both of which are subject to range restriction. However, although a correlation coefficient (parametric or otherwise) does not measure the degree of absolute

agreement (i.e., a perfect correlation may exist when absolute agreement does not exist),[10] some versions of analysis of variance arguably measure both rater variance and agreement.[11] Moreover, at the nominal level the debate regarding reliability and agreement is superfluous. Here, rating categories do not differ quantitatively but qualitatively (i.e., categorically), and so reliability, consistency, or agreement becomes an absolute, and proportionality is not applicable (Tinsley & Weiss, p. 361).

The idea of consistency of ranking introduces yet another term—concordance. Concordance is special form of interrater agreement that is measured by non-parametric tests of correlation solely on the basis of absolute ranking or ordering. Fortunately, there appears to be no dispute regarding the conceptual meaning of this term, although terminology may differ, and so requires no further clarification.

Avoiding the debate are authors such as Rosenthal and Rosnow (1991), who prefer to define the concepts of reliability, consistency, and agreement, in "general terms" (p. 46). For example, reliability is simply defined as consistency or stability, is measured quantitatively, and is reported in terms of a reliability coefficient. The reliability coefficient is defined as a "generic name for the degree to which what is measured is relatively free from measurement fluctuations." Moreover, Rosenthal and Rosnow define observer agreement as a reliability index that applies to observations made by different raters at the same time, and define consistency as a reliability index that applies to observations made by the same rater at different times (i.e., stability). Similarly, other researchers define between-observer reliability as a reliability index that applies to

[10] For a computational example, see Tinsley and Weiss, 1975.
[11] This point will be discussed in more detail in the section on intraclass correlation, under the heading "Decision #2."

observations made by different raters at the same time and within-observer reliability or within-observer consistency as a reliability index that applies to observations made by the same rater at different times (e.g., Martin & Bateson, 1993, p. 117).

Although one might avoid the debate, confusion of terms may lead to confusion of methodology, which in turn leads to confusion regarding the appropriate technique for estimating reliability within a particular context (Fleenor et al., 1996, p. 368). For example, Kozlowski and Hattrup (1992) cite recent instances in which researchers still confuse agreement with consistency and state that there is little consensus on the appropriate use of the numerous indices that have been proposed (p. 162).

Definitions

For the purposes of this study, although interrater consistency and agreement may arguably be conceptually distinct, they are still both consistent with the general notion of reliability. However, to distinguish between consistency and agreement is still useful, not for the purposes of debating whether one or both is compatible with classical reliability theory, but for choosing the appropriate method in the circumstances. As a result, similar to Rosenthal and Rosnow (1991), reliability is defined generally; i.e. the degree to which whatever is being measured is free from measurement error (p. 46).

Agreement is defined in absolute terms (i.e., essentially absolute consensus on qualitative or quantitative data), and is the only method by which nominal or categorical data can be assessed in terms of reliability. Agreement, though, is not restricted to categorical data, and may be redefined quantitatively when applied to ordinal data or data with true numeric properties.

Consistency is defined as being correlational in nature (as defined by classical reliability theory), and is the proportionate consistency of variance among raters. Therefore, as an index of reliability, consistency is measured by parametric tests, where corresponding coefficients indicate the extent to which the relative judgments of raters agree. Thus, consistency is inapplicable to categorical data.

Concordance is special form of interrater agreement that refers to consensus on ranking or ordering, and it is measured by non-parametric tests of correlation.

Alternate Theories

This study focuses on classical test theory, which dominates the literature on measurement theory. According to Cronbach (1984), classical test theory "is the basis for most reports on error found in test manuals" (p. 161). However, previously mentioned alternate theories have been proposed, such as generalizability theory and item response theory. Cronbach (1984), for example, argues that generalizability (G) theory is a more flexible procedure for research on assessment, and that analysis of error components by means of G theory "tells more about a measuring procedure than the traditional analysis" (p. 161).

In contrast to classical test theory's one true score, G theory refers to a domain or universe, which is a set of observations or scores of interest. From a sample of behavior, such as an individual's typing scores, a teacher may wish to generalize to a particular domain of interest, such as an individual's average typing scores (Cronbach, 1984, p. 161). Notably, a universe score may be similar to a true score. The difference, however, is that classical test theory interprets error variance as random chance only and assumes one "true score." In contrast, G theory interprets observed scores as an estimate of true

scores for a particular universe, assuming that all test scores are available (Rosenthal & Rosnow, 1991, p. 50).

Generalizability theory interprets error more liberally, recognizing "alternate universes of generalization" (Cronbach, 1984, p. 161). The flexibility of G theory, then, lies in the researcher's ability to choose the universe, depending upon the purpose of the measurement. Therefore, according to Cronbach, G theory "clarifies results from classical reliability formulas" by being more specific regarding exactly what counts as error (p. 161). Moreover, Cronbach argues that G theory facilitates the design of new and innovative measuring procedures.

It appears that G theory is not in conflict with classical test theory; rather, it is an alternative (James et al., 1993, p. 307; Traub, 1994, p. 3) that uses classical reliability formulae (Cronbach, 1984, p. 163). In other words, reliability is studied within the context of generalizability theory (Li et al., 1996). The emphasis of G theory is not on one true score with its restrictive interpretation of error variance; i.e., classical test theory assumes that for the dimension of interest an individual has a true score that will not change with repeated applications of the same test (Kaplan & Saccuzzo, 1982, p. 89). Instead, G theory emphasizes an analysis of error components within any number of universes, depending upon the particular purpose of the researcher.

In conclusion, while this study has only briefly commented on generalizability theory and has not commented on item response theory, the purpose here was simply to recognize that alternatives to classical test theory exist. Despite the merit of pursuing this subject further, for the purposes of this study classical test theory provides a sound theoretical basis to assess interrater reliability in an assessment center.

Interrater Reliability Tests

Test reliability is generally concerned with the measurement of temporal stability (i.e., test-retest reliability) and/or internal consistency (i.e., the reliability of test items). An assessment center does not normally measure temporal stability—assessing a group of candidates and re-testing the same group at a later time (also known as the rate-rerate method). Not that the concept is not valid, but the time and expense necessary to conduct an assessment center preclude this type of reliability measurement.

Besides, factors such as memory effect confound the variables of real change and test-retest reliability. For example, if a difference exists between the reliability index of one assessment administration and a subsequent administration, an issue arises regarding whether the difference is due to real change in candidate behavior, artificial change due to candidates remembering the exercises, change in rater assessment due to error or bias, or any combination of these factors (Kaplan & Saccuzzo, 1982, p. 93; Tinsley & Weiss, 1975, pp. 361, 373).

As a result, most assessment center administrators focus on internal consistency; that is, consistent rater assessments (analogous to items in a traditional written test) for one group of candidates. This is sometimes described as the replication of rater ratings, and is often measured by self-correlation techniques (Dick & Hagerty, 1971, p. 152; Guilford & Fruchter, 1978, p. 411; Tinsley & Weiss, 1975, p. 361). For numeric data, there are a number of self-correlation techniques, including well known parametric tests such as the product-moment correlation (Pearson's *r*), the intraclass correlation (ICC), and Cronbach's (1951) alpha (α).[12]

[12] As previously discussed, there are also alternative methods that are not correlational in nature, such as Tinsley and Weiss' (1975) T-index, which is a measure of agreement.

For non-numeric data, there are non-parametric tests such as the correlational approach based on absolute ranking, Cohen's (1960) *kappa*, Finn's *r*, and the "phi" coefficient (see Cramer, 1997, pp. 333-335). In general, non-parametric tests may be applied to any level of data, but the result is usually a loss of statistical power and so is not recommended for numeric data (Vito & Latessa, 1989). Consequently, this section focuses on providing a detailed explanation of the parametric indices most often used in assessment centers for measuring interrater consistency.

Parametric Tests

There are four basic levels of measurement (i.e., scales that yield a numerical representation of the values of a variable of interest): nominal, ordinal, interval, and ratio. These scales yield values or numbers that convey very different properties and so determine how the numbers can be manipulated and interpreted (see Traub, 1994, p. 2, for the history of these terms; and Kachigan, 1991, pp. 10-15, for a more detailed discussion). [13] Nominal data have no true numeric properties (they cannot be added, subtracted, multiplied or divided) and can be described as categorical data that are classified along qualitative dimensions of interest (e.g., gender). Consequently, nominal data are appropriate for analyses by non-parametric methods only (non-parametric and parametric concepts are discussed below).

Similarly, ordinal data, strictly speaking, have no true numeric properties. Ordinal data assume the properties of nominal data, along with the properties of ordering or

[13] A variable is a characteristic (e.g., intelligence) of an object of interest (e.g., person). A variable differs in some substantive way, and can be measured operationally by means of a scale (e.g., intelligence test). Variables can be classified as qualitative (measured at the nominal or ordinal level) or quantitative (measured at the interval or ratio level). Moreover, variables may be classified as discrete (may only take on a finite number of values) or continuous (may take on any number of values along a continuum).

ranking along a continuum, although not in a relative sense (i.e., intervals between measurement points are not necessarily equal) (Tinsley & Weiss, 1975, p. 360). Theoretically, mathematical operations for both nominal and ordinal data are meaningless.

On the other hand, interval and ratio data have true numeric properties and are appropriate for analyses by both parametric and non-parametric methods. Interval data assume the properties of nominal and ordinal data, along with equal intervals between measurement points on the scale. For interval data, in the absence of an absolute zero point, only the mathematical operations of addition and subtraction are valid. Ratio data assume the properties of nominal, ordinal and interval data, along with an absolute zero that allows for the creation of ratios (Tinsley & Weiss, 1975, pp. 360-361; Miller & Whitehead, 1996, pp. 25-27). As a result ratio data may be added, subtracted, multiplied and divided.

The distinction between nominal, ordinal, interval, and ratio data is important because data type determines the appropriateness of parametric or non-parametric techniques. Briefly, in the field of statistics, population values are known as parameters and the statistical tests used to measure them are called parametric (Siegel, 1956, p. 2).[14] Parametric tests make many assumptions; for example, that data have true numeric properties (i.e., interval or ratio), that scores were drawn from a normally distributed population, and that scores were drawn from populations with equal variances.

[14] A parameter is a measurement of a dimension of interest (characteristic or variable) of a population (i.e., universe of interest), and a statistic is a measurement of a dimension of interest of a sample or subset of the population of interest.

Non-parametric techniques make fewer assumptions about parameters (Siegel, 1956, p. 3). For example, non-parametric tests do not assume a normal distribution or equal variances. Moreover, many non-parametric tests do not assume true numeric properties, but rather focus on absolute ordering or ranking of scores, and even classification of nominal (categorical) data when ordering is not possible. Notably, non-parametric tests measure absolute rank or order, while parametric techniques measure relative rank or order and so are dependent upon computation of means and variances. The issue of means introduces a significant difference between non-parametric and parametric tests; specifically, non-parametric tests focus on the difference between the medians of samples, while parametric tests focus on the difference between the means of samples. The computation of means requires data to have true numeric properties, while the computation of medians requires counts only (Siegel, p. 3).

Notwithstanding, parametric tests generally require true numeric properties for the dependent variable only, and most social scientists are not concerned with the differences between interval and ratio levels of measurement (Erickson & Nosanchuk, 1977, p. 167; Miller & Whitehead, 1996, pp. 27, 295). Clearly, nominal data require non-parametric tests, but the issue is not quite so clear concerning the statistical treatment of ordinal data because the distinction between ordinal and interval data is often blurred. In the social sciences, research data are frequently at the ordinal level where measurement intervals are logically equal, such as those found in Likert-type scales used to measure opinions and behavioral rating scales used at assessment centers (Tinsley & Weiss, 1975, p. 360). For example, Table 3-1, which reproduces the rating scale used by the Police Academy

assessment center, illustrates how ordinal level measurements can be interpreted as interval data.

Table 3-1

Police Academy assessment center rating scale

Rating	Score	Converted Score*
Excellent	5	55
Superior ability	5-	52
A great deal of ability	4+	48
Well above average[+]	4	45
Above average	4-	42
Slightly above average	3+	38
Average (competent)	3	35
Slightly below average	3-	32
Below average	2+	28
Well below average	2	25
Very little ability	2-	22
Poor	1+	18
Very poor	1	15

* Rather than converting base scores to decimals (e.g., 3+ to 3.33), the Police Academy uses the "converted score."

[+] The term "average" should be read as "competent."

In such cases, research has shown that if the assumption of equal intervals is not too severely violated, parametric tests can be meaningfully applied to ordinal data (Erickson & Nosanchuk, 1977, p. 167; Kachigan, 1991, pp. 13-14; Miller & Whitehead, 1996, pp. 272, 386; Tinsley & Weiss, 1975, pp. 360-361, citing Baker, Hardyck & Petrinovich, 1966). Thus, the assessment center administrator may use parametric tests on "quasi-interval" rater assessments, a standard practice that generally receives little attention in the literature.

The Correlational Approach

Interrater reliability is conceptually similar to test reliability—just as test or scale reliability may refer to consistency between items on a particular dimension, interrater reliability may refer to consistency between raters on a particular dimension (e.g., in the

case of assessment centers, the overall suitability of a candidate for a particular job). Therefore, some traditional test reliability procedures are appropriate for use with judges or raters (e.g., Cronbach's (1951) alpha coefficient and the Spearman-Brown correction formula).

In this discussion of reliability, the correlational approach is discussed first. The product-moment correlation[15] indicates the strength of the association or relationship between two variables by means of a ratio. This ratio results in a coefficient that explains the strength of the relationship in terms of a percentage (Miller & Whitehead, 1996, p. 322-323). As noted in the section on theory, a reliability coefficient is really a function of the correlation of one measure x with another measure x_1 (Cronbach, 1984, p. 160; Shrout and Fleiss, 1979, p. 422), although some authors criticize the traditional correlational method (Pearson's r) for estimating interrater reliability (e.g., Fleenor et al., 1996, p. 378; Tinsley & Weiss, 1975, p. 366).[16]

The correlational approach gives the mean reliability of a single rater and therefore underestimates the mean reliability of a group of raters (Fleenor et al., 1996, p. 373; Rosenthal & Rosnow, 1991, p. 51; Tinsley & Weiss, 1975, p. 364). For example, if a correlation coefficient of .7 is calculated for two raters rating four candidates at an assessment center, the coefficient represents the reliability of either single rater. If the composite reliability (also known as the aggregate reliability or effective reliability) of

[15] "Moments" refer to standardized distances from the mean for variables X and Y, which are multiplied by each other to form "products." Hence the term "product-moment correlation" (rf. Kachigan, 1991, pp. 130; Rosnow & Rosenthal, 1996, p. 236; and Traub, 1994, p. 11).

[16] Criticisms are that the correlational approach (1) assumes equal variances and so probably overestimates R, (2) doesn't consider differences between rater means as error, and (3) doesn't partition error as does ANOVA.

both raters is of interest, the estimate for a single rater must be corrected upwards by the Spearman-Brown formula (described in the discussion on theory):

$$R = \frac{n\bar{r}}{1 + (n-1)\bar{r}},$$

where R is the effective or corrected reliability coefficient, n is the number of raters, and $\bar{r}$ is the mean correlation among all raters (Rosenthal & Rosnow, pp. 51-52; Tinsley & Weiss, p. 365). If this formula is applied to the example above, the corrected reliability is .82 compared to the mean reliability .7 of a single rater:

$$R = \frac{(2).7}{1 + (2 - 1).7} = .82.$$

As noted in the section on theory, a coefficient of 0.0 indicates no reliability while a coefficient of 1.0 indicates perfect reliability.[17]

A real example can be provided be examining class 245a at the Police Academy assessment center, which generally uses three or four raters to assess five to six candidates on a 13-point rating scale (as shown in Table 3-1) in order to identify the most suitable candidates for the occupation of police constable. The data for this class are shown in Part I of Table 3-2, where correlations were calculated for each pair of raters: A with B, A with C, and B with C. Subsequently, as shown in Part II, a mean correlation of .44 was calculated ((.136 + .645 + .535)/3).

[17] Actually, a correlation coefficient may range from –1.0 to +1.0, but a negative correlation coefficient between two raters would indicate that their assessments were inversely related, and is inconsistent with classical reliability theory, where R is bounded by 0.0 and 1.0.

Using the Spearman-Brown formula, the corrected reliability was calculated as follows:

$$R = \frac{(3).44}{1+(3-1).44} = \frac{1.32}{1.88} = .70.$$

Table 3-2[18]

Reliability for class 245a: Correlational approach

Part I: Rater Scores

	Raters		
Candidates	A	B	C
1	38.208	40.375	39.333
2	37.333	39.500	38.167
3	34.700	40.167	35.333
4	37.778	40.333	37.667
5	37.545	41.375	39.778
6	36.636	40.792	40.500

Part II: Rater correlations

Raters	Correlation *r*
AB	.136
AC	.645
BC	.535
Mean	.439

The corrected reliability can also be quickly estimated by using a chart developed by Rosenthal and Rosnow (1991, p. 53, citing Rosenthal, 1987), which gives the effective reliability as .71. This chart, based on the Spearman-Brown formula, provides the researcher with a quick method to estimate an *R* value by locating the intersection between the *n* (the number of raters) row and the $\bar{r}$ (mean reliability) column that coincides with the researcher's data. The chart is also useful as the researcher can

[18] This data were taken from a Police Academy assessment center held in 1998.

quickly estimate the total number of raters necessary to obtain a desired *R* value based on a known mean reliability for a known number of raters; or alternatively, estimate the mean reliability of a single rater based on a known *R* value for a known number of raters.[19]

Note that the value $\bar{r}$ was calculated by averaging the correlation coefficients according to normal arithmetical rules, which is technically incorrect because correlation coefficients have properties that differ from ordinary numbers (Kaplan & Saccuzzo, 1982, p. 95; Martin & Bateson, 1993, pp. 140-141). To find the mean of any number of correlation coefficients (i.e., Pearson's *r*), each correlation coefficient should first be transformed by way of Fisher's *Z* transformation formula. The mean of these *Z* transformations may be calculated according to normal arithmetical rules, and subsequently converted back into a correlation coefficient.[20]

It is this value of $\bar{r}$ that represents the true mean for correlation coefficients, and is especially useful for determining the difference between independent correlation coefficients. However, for the purposes of calculating interrater reliability coefficients, violation of this rule is considered acceptable practice (Rosenthal & Rosnow, 1991, pp. 51-54; 431-432), and is usually not mentioned in the literature except for the occasional footnote (e.g., Kaplan & Saccuzzo, 1982, p. 95). For example, from the data in Table 3-2, $\bar{r}$ was calculated as .44, while the true $\bar{r}$ calculated by using Fisher's *Z* transformation formula is approximately .46. As shown in this example, if there are small differences

[19] The Spearman-Brown correction formula can be altered into what is known as the Spearman-Brown prophecy formula to calculate the same results (rf. Kaplan & Saccuzzo, 1982, p. 107).
[20] Tables for such *Z* transformations are available in texts written by Snedecor and Cochran (1980), reprinted in Rosenthal and Rosnow (1991), and Fisher (1932) reprinted in Guilford and Fruchter (1978).

between the *r*s that are to be averaged and the *r*s are not large, then the estimated mean $\bar{r}$ will be close to the true $\bar{r}$. If these conditions are not badly violated "a simple arithmetic mean will suffice" (Guilford & Fruchter, 1978, p. 330), otherwise transformations are easily calculated from *Z* tables.

The Intraclass Approach

In addition to the correlational approach, Rosenthal and Rosnow (1991) recommend analysis of variance as an excellent method to estimate interrater reliability (pp. 55-56; see also Tinsley & Weiss, 1975, p. 373).[21] The use of analysis of variance (ANOVA) to estimate reliability was pioneered by Jackson (1939), Hoyt (1941), and Alexander (1947), and today is probably the most widely used method for estimating interrater reliability on ordinal or interval data (Cramer, 1997, p. 333; Dick & Hagerty, 1971, p. 149; Fleenor et al., 1996, p. 373; Kozlowski & Hattrup, 1992, p. 162-163; Rosenthal & Rosnow, 1991, p. 50; Shrout & Fleiss, 1979, p. 420; Tinsley & Weiss, 1975, pp. 359, 363; Whitehurst, 1985, p. 458).[22]

Based on the data in Table 3-2, the mean and corrected reliability coefficients calculated by the correlational approach were .44 and .70 respectively. Based on the same data, but using formulae developed to estimate interrater reliability from ANOVA calculations, as shown in Table 3-3 and discussed below, the interrater reliability for a single rater is .41 and the corrected reliability is .68.

---

[21] Note that analysis of variance is also appropriate for two or more raters, as it will test for differences between two or more means (rf. Koosis, 1997, p. 175; Rosenthal & Rosnow, 1991, p. 317; Rosnow & Rosenthal, 1996, p. 283).

[22] According to Dick and Hagerty, Hoyt's method became the most widely used. He showed that the Kuder-Richardson (1937) formula 20 (applicable to dichotomously scored test items) was identical to his method.

Table 3-3

ANOVA table (repeated-measures) (from Table 3-2)

Variance Source	SS	*df*	MS
Between subjects	16.38	5	3.276
Within subjects			
Between raters (treatments)	34.76	2	17.380
Error & interaction (residual)	10.65	10	1.065
Total	61.79	17	3.635

Using the "intraclass" approach, the formula[23] for estimating the mean reliability of an individual rater is:

$$\bar{r} = \frac{MS\ \text{subjects} - MS\ \text{error}}{MS\ \text{subjects} + (k-1)(MS\ \text{error})},$$

where the means squares (*MS*) are given in the ANOVA table, and k is the total number of raters. Therefore:

$$\bar{r} = \frac{3.276 - 1.065}{3.276 + [(3-1)1.065]} = .41.$$

The formula for estimating the corrected reliability is:

$$R = \frac{MS\ \text{subjects} - MS\ \text{error}}{MS\ \text{subjects}}.$$

Therefore:

$$R = \frac{3.276 - 1.065}{3.276} = .68.$$

[23] This formula and the one following are the two most commonly used versions of the intraclass correlation coefficient (ICC) that are found in various texts and articles (e.g., Guilford & Fruchter, 1978, p. 270; Rosenthal & Rosnow, 1991, pp. 55-56, 432; Shrout & Fleiss, 1979, pp. 423, 426). The ICC method and its various configurations will be discussed in greater detail later in this study.

Comparing the results of the correlational approach with those of the intraclass method indicates that the differences between them are quite small (see also Table 3-9), similar to results found by Cronbach (1984, pp. 169-170), Fleenor et al. (1996, p. 377-378), and Rosenthal and Rosnow (1991, pp. 55-56). As is the case generally, 0.0 indicates no reliability while 1.0 indicates perfect reliability (Tinsley & Weiss, 1975, p. 363).

The use of analysis of variance for estimating interrater reliability is known as the intraclass method, and yields what is commonly called an intraclass correlation coefficient (ICC), for either *r* or *R* (Cronbach, 1984, p. 169; Rosenthal & Rosnow, 1991, pp. 50, 55-56, 430-432).[24] Defining this coefficient, Shrout and Fleiss (1979) state that "the ICC is the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target" (p. 422). Therefore, according to Shrout and Fleiss, the ICC is a bona fide correlation coefficient that is often identical to the sum of true target variance divided by the sum of target and error variance obtained by standard analysis of variance calculations (see also Fleenor et al., 1996, pp. 373-374). The ICC is only delineated from the Pearson' *r* in that it uses the computational procedures of analysis of variance (Tinsley & Weiss, 1975, p. 363). Hence, the ICC indicates the degree of association of observations made by raters on a target, such as subjects or candidates at an assessment center, and is a "good estimate of

[24] The ICC has also been known as an alpha (α_K) coefficient, this Greek symbol representing the reliability coefficient for a composite of *k* raters (Cronbach, 1984, p. 169). Today, however, the term alpha coefficient and the symbol (α) as it applies to reliability is more often associated to Cronbach's (1951) formula described in his influential article, "Coefficient Alpha and the Internal Structure of Tests." Notably, Cronbach (1984) acknowledges that the term coefficient alpha is conventionally linked to his formula (p. 170). Interrater reliability can also be estimated with Cronbach's alpha, which will be discussed later.

the mean correlation obtained by correlating all possible pairs of observations made on subjects" (Rosenthal & Rosnow, 1991, p. 431).

In addition to the convenience of using ANOVA techniques when there are numerous raters and subjects, ICC is also recommended over the correlational approach because it identifies the variance associated with each component (Tinsley & Weiss, 1975, p. 366, citing Ebel, 1951). When ICC is used to calculate interrater reliability, the between mean squares (subjects) component estimates the group score variance, and so the square root of this component would estimate the standard deviation for group scores. Approximately two-thirds of subject scores will fall within one standard deviation of the mean group score. The within mean squares component is partitioned by repeated-measures ANOVA into conditions variance (i.e., how much group averages fluctuate from one rater to another), and residual variance. The residual component includes any kind of unsystematic error associated to individual differences observed under each condition (i.e., rater) (Cronbach, 1984, pp. 166-168), including interaction between rater and subject (Guilford & Fruchter, 1978, pp. 268-269; Rosenthal & Rosnow, 1991, p. 396), although of course all error is assumed to be random.

ANOVA: Two-way vs. Repeated-measures. There can be some confusion regarding which ANOVA technique is appropriate for calculating an ICC. Cronbach (1984), Shrout and Fleiss (1979), and Tinsley and Weiss (1975), for example, describe two-way ANOVA procedures while Rosenthal and Rosnow (1991) describe one-way repeated-measures ANOVA procedures for calculating ICC. According to Fleenor et al. (1996), Winer (1971, p. 283) "developed a single factor repeated-measures analysis of

variance to calculate the reliability of the mean *k* of raters" (p. 374). Both techniques appear to be equally used, and both yield the same results.

Two-way ANOVA without replications is a special form of two-way ANOVA that can be applied to interrater reliability. It includes two factors of interest (subjects and raters); but because both are not divided into levels, interaction is not partitioned from error (Cramer, 1996, pp. 338-339; Guilford & Fruchter, 1978, p. 268). Repeated-measures ANOVA is conceptually more appealing because of its fit with classical reliability theory—by repeatedly giving the same test to the same person, the expected score should approximate the true score (Traub, 1994, p. 21). Operationally, the repeated-measures design facilitates the observation of two or more treatments (e.g., assessments by raters) on the same sampling units (e.g., subjects). Here, subjects are "crossed" by treatment conditions (known as a within subject, matched samples, or repeated-measures design) rather than being "nested" within them (a between subject design) (Miller & Whitehead, 1996, pp. 291-292; Rosenthal & Rosnow, 1991, p. 392). Thus, the repeated-measures method more accurately describes the ANOVA procedure used for ICC (Fleenor et al., 1996, p. 375).

Versions of the Intraclass Correlation Coefficient. There are six different forms or versions of ICC used to calculate a reliability coefficient in which *n* subjects are assessed by *k* raters, the choice of version depending upon the rater model and the purpose of the reliability study (Cramer, 1997, pp. 336-337; Shrout & Fleiss, 1979, p. 420; Tinsley & Weiss, 1975, p. 363). That the choice of version is properly informed is underscored by the fact that the different versions can yield very different results when applied to identical data. However, according to Shrout and Fleiss, "many researchers

are not aware of the differences between the forms, and those who are often fail to report which form they used" (p. 420). Consistent with the findings of Shrout and Fleiss, this research found that most texts and articles are incomplete in their description of intraclass models, including errors in older references, all of which unnecessarily complicates the task of either the researcher or the assessment center administrator.

The article on the intraclass method by Shrout and Fleiss (1979) is especially notable because it clarifies much of the methodological confusion surrounding its use. The six different versions of ICC are described in detail, along with the conditions under which the use of each version is appropriate. First, Shrout and Fleiss describe three basic decisions that the researcher must make (p. 420):

Decision #1: Deciding on appropriate ANOVA technique.

Decision #2: Deciding if differences between raters' means squares are relevant (i.e., whether to include them in the error term).

Decision #3: Deciding if the unit of analysis is an individual rater or the mean of several raters.

Decisions one and two address the issue of choosing the appropriate statistical technique, while decisions two and three address the purpose of the reliability study. In a typical assessment center, each subject is assessed independently by two or more raters, which is typical for interrater models in general. Within this model, there may occur three separate configurations (Shrout & Fleiss, 1979, pp. 420-421):

Configuration #1: Each subject in a group is rated by an unknown set of *k* raters, selected randomly from a pool or population of raters.

Configuration #2: Each subject in a group is rated by the same set of unknown *k* raters, selected randomly from a pool or population of raters.

Configuration #3: Each subject in a group is rated by the same set of known *k* raters, who are the only raters of interest.

Each configuration, because of assumptions associated with random raters and the ability to generalize, requires its own mathematical technique, all corresponding to standard ANOVA methods, described as follows (Shrout & Fleiss, 1979, pp. 421-423; see also SPSS® Base 10.0, 1999, p. 367):

Configuration #1: One-way ANOVA (random effects model).

Configuration #2: Repeated-measures ANOVA (random effects model).

Configuration #3: Repeated-measures ANOVA (mixed effects model).

Random effects models, in which it is assumed that raters are selected randomly from a population, permit the results to be generalized to other raters (it is assumed that the subjects in all configurations are selected randomly from a larger pool of subjects). Because the raters are fixed in configuration three, the mixed effects design is the appropriate model (Shrout & Fleiss, 1979, pp. 421-422; see also Rosenthal & Rosnow, 1991, pp, 395-398; and SPSS® Base 10.0, 1999, pp. 367-368). Note that the SPPS® Base 10.0 computer program calculates configurations two and three with two-way ANOVA models, but as previously noted, the results are identical to one-way repeated-measures ANOVA.

Decision #1: Appropriate ANOVA method. The selection of one-way ANOVA or repeated-measures ANOVA depends upon whether configuration one is applicable, or whether configuration two or three is applicable. For configuration one, one-way

ANOVA, which yields a between subjects mean square (BMS_S) and a within subjects mean square (WMS_S), is appropriate (Cramer, 1997, pp. 337-338; Shrout & Fleiss, 1979, p. 422; Tinsley & Weiss, 1975, p. 364). The within subject mean square (WMS_S) includes both the between raters mean square and the error term (Tinsley & Weiss, p. 364). The following formula (Shrout & Fleiss, 1979, p. 423), then, estimates the population value ρ for configuration one:

$$\text{ICC}(1,\ 1) = \frac{\text{BMS}_S - \text{WMS}_S}{\text{BMS}_S + (k-1)\text{WMS}_S},$$

where ICC indicates the intraclass correlation coefficient, the numbers in the parenthesis refer to the configuration and an individual rater ($\bar{r}$) respectively, the subscript "$_S$" refers to subjects, and "k" is the number of raters.

To calculate one-way ANOVA, one can either use traditional calculations (for example, see Miller & Whitehead, 1996, pp. 285-287) or use the calculations from repeated-measures ANOVA. If one-way ANOVA is calculated traditionally from a sum of squares table as described by Miller and Whitehead (1996), one must be careful to arrange the table so that between groups mean squares apply to the subjects (vis-a-vis between raters). Specifically, the subjects variable is ordered horizontally in the table, while the raters variable is ordered vertically.[25]

Notwithstanding that one-way ANOVA is the appropriate method, the data from one-way repeated-measures ANOVA can be used, where the between raters sum of squares and the residual sum of squares are added, as are their respective degrees of

[25] For two-way or repeated-measure ANOVA, how the table is ordered does not matter because the rows or columns can apply either to the raters or the subjects.

freedom, to form the within subjects sum of squares component. For example, in Table 3-3 above, the between raters sum of squares (34.76) would be added to the residual sum of squares (10.65), for a total of 45.41, which in one-way ANOVA is the within subjects sum of squares component. This sum of squares can then be divided by the associated degrees of freedom (2 + 10 = 12) to yield the within subjects mean square (WMS_S) (45.41 ÷ 12 = 3.78), as indicated in Table 3-4 below (compare with Table 3-3). As a result, one-way ANOVA calculations are unnecessary, as repeated-measures ANOVA will provide all information necessary to calculate any ICC version.

Table 3-4

ANOVA table (one-way) (calculated from Table 3-2)

Variance Source	SS	*df*	MS
Between subjects	16.38	5	3.276
Within subjects	45.41	12	3.784
Total	61.79	17	3.635

For configuration two or three, repeated-measures ANOVA is appropriate (Shrout & Fleiss, 1979, p. 423; Tinsley & Weiss, 1975, p. 363; see also SPSS® Base 10.0, 1999, p. 367). As previously noted, repeated-measures ANOVA partitions the within subject mean squares (WMS_S) into between raters mean squares (BMS_R) and the residual (error + rater x subject interaction) mean squares (MS_E). Although ANOVA calculations for configurations two and three are the same, configuration two assumes that raters are randomly selected (which allows one to generalize—similar to configuration one—to any set of the same size of known *k* raters), while configuration three assumes that the known raters are fixed (who are then the only raters of interest) (Shrout & Fleiss, 1979, pp. 423, 427).

For configuration two, Shrout and Fleiss (1979, p. 423), citing Rajaratnam (1960) and Bartko (1966), provide the following ICC formula, which estimates the population value ρ as follows:

$$\text{ICC}(2,\ 1) = \frac{\text{BMS}_S - \text{MS}_E}{\text{BMS}_S + [(k-1)\text{MS}_E] + [k(\text{BMS}_R - \text{MS}_E)/n]}.$$

And for configuration three, Shrout and Fleiss (p. 423) provide the following ICC formula, which estimates the reliability for a fixed group of known raters (see also Guilford & Fruchter, 1978, p. 270; Rosenthal & Rosnow, 1991, pp. 56, 432; Tinsley & Weiss, 1975, p. 364):[26]

$$\text{ICC}(3,\ 1) = \frac{\text{BMS}_S - \text{MS}_E}{\text{BMS}_S + (k-1)\text{MS}_E}.$$

[26] Note that although ANOVA calculations are the same for configurations two and three, where the proper error term for between raters' mean squares is the same for both random and mixed effects models (Rosenthal & Rosnow, 1991, p. 398), different formulae as noted here produce different reliability coefficients for the reasons noted. However, SPSS® Base 10.0 (1999) will produce identical reliability coefficients for both the two-way random and mixed effects models (see SPSS® Base 10.0 manual, pp. 367-368). For example, if the hypothetical data from Shrout and Fleiss (1972, p. 423) are entered into SPSS® Base 10.0, analysis results in identical reliability coefficients of .71 for configurations two (random) and three (mixed). Alternately, Shrout and Fleiss, by applying two-way (without replications) ANOVA calculations to the appropriate formula, calculate reliability coefficients of .21 and .71 for configurations two and three respectively (p. 424). Because of the assumption of generalizing in configuration two, it seems logical that Shrout and Fleiss, who are cited extensively in the literature, are correct (but the difference between approaches is due to a theoretical interpretation). The conflict aside, due to discrimination liability the focus of an operational assessment center is usually fixed, where configuration three is the appropriate choice. As a result, SPSS® Base 10.0 is a useful program for assessment center administrators should they choose to use the intraclass method. Notably, this is a good example of the confusion surrounding interrater reliability discussed in the introduction to this chapter; hopefully, this explanation will clarify any apparently inconsistent results between SPSS® Base 10.0 and ANOVA calculations applied to the formulae provided by Shrout and Fleiss.

Decision #2: Relevancy of between raters' mean squares. This issue was introduced in the discussion above regarding decision one—the methodological choice between configurations one and two in which raters are random and configuration three in which the raters are fixed. In configurations one and two, because raters are random, with the underlying assumption that interrater reliability results are generalizable to a larger population of raters, the between raters' mean squares component is included in the formulae. On the other hand, in configuration three where the focus is on one rater or a fixed set of *k* raters, which precludes an assumption to generalize beyond the rater(s) under study, the between raters' mean squares component is not included in the formula (Shrout & Fleiss, 1979, p. 424; Tinsley & Weiss, 1975, pp. 363-364).

Notably, returning to the debate concerning agreement and consistency within the concept of interrater reliability, Shrout and Fleiss (1979) reason that when rater variance is included within the error component (for configurations one and two), the ICC index can be interpreted as rater "agreement" (p. 425; see also Cramer, 1997, pp. 337-338; and Tinsley & Weiss, 1975, pp. 363-364). Therefore, formulae for agreement may be found in configurations one and two for either the mean correlation of one rater (see the discussion above on decision one), or the corrected correlation of *k* raters (see the discussion below on decision three). It can be argued, then, that the issues of generalizability and agreement within the context of interrater reliability are connected. When configuration one or two is used, Shrout and Fleiss (1979) postulate that the raters are interchangeable because the raters were selected randomly and between raters

variance is included in the error term.[27] Alternately, when rater variance is ignored, as in configuration three, the ICC index can be interpreted as the consistency of a fixed set of raters (Shrout & Fleiss, p. 425).[28]

It should be noted here that Tinsley and Weiss (1975) argue that generalizations are not desirable within the context of interrater reliability because "interrater reliability and agreement are functions of the subjects rated, the rating scales used, and the judges making the ratings" (p. 373). They conclude that measures of interrater reliability are best used as one trial estimates, although they do not dismiss the notion of generalizability when between raters variance is included in the error term (p. 364).

Decision #3: Individual rater or the mean of several raters. The ICC versions previously discussed yield the expected reliability coefficient $\bar{r}$ (i.e., the mean correlation of a single rater). In the situation where the ratings of k number of raters is of interest, the expected reliability coefficient R "will always be greater in magnitude than the reliability of the individual rating, provided the latter is positive" (Shrout & Fleiss, 1979, p. 426, citing Lord & Novick, 1968). As previously discussed in the correlational approach section, the ratings of k number of raters are known as the aggregate reliability or effective reliability, which is the composite or corrected reliability of all raters who rate a particular target (Rosenthal & Rosnow, 1991, p. 51).[29]

---

[27] For further discussions on agreement, generalizability, and interchangeability, see Fleenor et al. (1996, p. 368), Kozlowski and Hattrup (1992, p. 163), and Tinsley and Weiss (1975, p. 364), citing Ebel (1951) and Bartko (1966).

[28] Interestingly, in 1976 Bartko argued that consistency was not an appropriate reliability concept for raters, but this argument was not supported in the literature. For example, Shrout and Fleiss (1979) rejected his conclusions, calling them unwarranted and misleading (p. 425).

[29] Notably, the correction is based on the Spearman-Brown correction formula, which was introduced in the section on history.

Corresponding to the formulae for one rater, Shrout and Fleiss (1979) describe the following formulae for k raters for configurations one, two, and three (p. 426). The formula (see also Tinsley & Weiss, 1975, p. 364) for k raters that corresponds to configuration one is:

$$\text{ICC}(1, k) = \frac{\text{BMS}_S - \text{WMS}_S}{\text{BMS}_S}.$$

The formula for k raters that corresponds to configuration two is:

$$\text{ICC}(2, k) = \frac{\text{BMS}_S - \text{MS}_E}{\text{BMS}_S + [(\text{BMS}_R - \text{MS}_E) / n]}.$$

And the formula (see also Tinsley & Weiss, p. 365, citing Ebel, 1951; Guilford & Fruchter, 1978, p. 270; and Rosenthal & Rosnow, 1991, p. 55) for k raters that corresponds to configuration three is:

$$\text{ICC}(3, k) = \frac{\text{BMS}_S - \text{MS}_E}{\text{BMS}_S}.$$

ICC versions most likely to encountered in texts and articles are $ICC_{3,1}$ and $ICC_{3,k}$, which were illustrated by way of example in the discussion on the data in Table 3-3 (see also Table 3-9). Note their similarity (especially $ICC_{3,k}$) to how the reliability ratio was conceptualized in the section on theory, error being partitioned from the observed score in order to estimate the true score (rf. Dick & Hagerty, 1971, p. 151). Because of

the frequency with which these two ICC versions are used, rather than being identified as $ICC_{3,1}$ and $ICC_{3,k}$, they are often identified by $\bar{r}$ and R only, as noted below:[30]

$$\bar{r} = \frac{BMS_S - MS_E}{BMS_S + (k-1)MS_E}, \text{ and}$$

$$R = \frac{BMS_S - MS_E}{BMS_S}.$$

Cronbach's Alpha

Cronbach's (1951) alpha (α) coefficient (an extension of the Kuder-Richardson formula 20)[31] is by far the most popular method for assessing the internal consistency of test items where the items are not scored dichotomously (Carmines & Zeller, 1979, p. 44). Cronbach's alpha assesses the extent to which different items on a test are measuring the same dimension; for example, a high alpha indicates test items are consistent with each other and are probably measuring the same trait (Kaplan & Saccuzzo, 1982, p. 103).

Because items of a test are analogous to raters in an assessment center, alpha is also a useful index of interrater reliability, and is appropriate with two or more raters (rf. Fleenor et al., 1996, p. 373).[32] Cronbach's alpha is also noteworthy within the context of

[30] Note that Tinsley and Weiss (1975, p. 364) wrote the formula for ICC (version 3, 1) with the symbol R that is described by Shrout and Fleiss (1979) as technically indicating ICC (version 3, k) (p. 364). To avoid confusion, when reading Tinsley and Weiss the reader should substitute the symbol $\bar{r}$ for this formula. Tinsley and Weiss use r to denote the Finn index, which may be used on ordinal data to assess average reliabilities of individual raters (p. 364).

[31] See Kaplan and Saccuzzo (1982, pp. 99-103) and Traub (1994, p. 87).

[32] Fleenor et al. state "three or more raters." However, alpha is appropriate for two or more raters just as it is appropriate for a test with two items, and just as ANOVA is appropriate for analysis of two or more groups (see Table 3-9 for an example).

this study because of its relationship with the intraclass method. According to Shrout and Fleiss (1979, p. 426) and Mehrens and Lehmann (1978, p. 99), the "ICC (3, *k*) is equivalent to Cronbach's (1951) alpha" and yields exactly the same coefficient (rf. Cronbach, 1984, pp. 169-170; see also Cramer, 1997, pp. 336, 341-343).[33]

For example, for the data in Table 3-2, reproduced in Table 3-5 below to show how alpha is calculated (discussed below), the corrected reliability *R* was estimated to be .68 by both $ICC_{3,k}$ and Cronbach's alpha. For intraclass *R*, when calculated by hand using the sum of squares method described by Miller and Whitehead (1996, pp. 292-295), *R* was .675; and when calculated by Aiken's (1996) software and SPSS® Base 10.0 (1999), *R* was .676 and .678 respectively. For Cronbach's alpha, when calculated by hand using the sum of squares method described by Ebel (1972, p. 420) or the shorter method described by Kaplan and Saccuzzo (1982, p. 103) and Cronbach (1984, p. 170), α was .678; and when calculated by SPSS® Base 10.0, α was .678.

Table 3-5

Reliability: Cronbach's alpha (calculated from Table 3-2)

	Raters			
Candidates	A	B	C	Total
1	38.208	40.375	39.333	117.916
2	37.333	39.500	38.167	115.000
3	34.700	40.167	35.333	110.200
4	37.778	40.333	37.667	115.778
5	37.545	41.375	39.778	118.698
6	36.636	40.792	40.500	117.928
mn ($\bar{x}$)	37.033	40.424	38.463	115.920
sd (σ)	1.147	.573	1.691	2.867
total score var ($\sigma^{2)}$	1.315	.328	2.859	8.220
total inter-item var (σ^2)			4.502	

[33] See also footnote 21, where Hoyt (1941) showed that the Kuder-Richardson (1937) formula 20 was identical to his method of using analysis of variance to estimate interrater reliability.

"True" alpha is based on covariance-variance (vis-a-vis "standardized item" alpha, which is discussed later), and can be written in terms of interrater reliability as follows (Carmines & Zeller, 1979, p. 44; Kaplan and Saccuzzo, 1982, p. 103):

$$\alpha = \mathrm{R} = \left(\frac{k}{k-1}\right)\left(\frac{S^2 - \sum Si^2}{S^2}\right), \text{ or alternately}$$

$$\alpha = \mathrm{R} = \left(\frac{k}{k-1}\right)\left[1 - \left(\frac{\sum Si^2}{S^2}\right)\right],$$

where R is the corrected reliability (Rosnow & Rosenthal, 1996, p. 128), k is the number of raters, S^2 is the variance of the sum of scores for each subject (i.e., total score variance), and Si^2 is the sum of the variance of the individual raters (i.e., total inter-item variance) (rf. Cronbach, 1984, p. 170). Using the data in Table 3-5, alpha was calculated as follows:

$$\alpha = \left(\frac{3}{3-1}\right)\left(\frac{8.22 - 4.502}{8.22}\right) = .678.$$

Fleenor et al. (1996) criticize the use of Cronbach's alpha for calculating interrater reliability by arguing that restriction of range may have the effect of decreasing variance in the inter-item scores, thus artificially increasing the obtained alpha value (p. 377). Fleenor et al. base their conclusion on the fact that "coefficient alpha is a function of the ratio of the sum of the inter-item covariances to the variance of the total score" (citing Ghiselli, Campbell & Zedeck, 1981; see also SPSS® Base 10.0, 1999, p. 362). For alpha to be greater than zero, the total test score variance must be greater than the sum of the variances for the individual items, and so the smaller the inter-item variance to the

total variance the larger the alpha (Kaplan & Saccuzzo, 1982, p. 100). However, the logic of Fleenor et al. is difficult to follow.

Covariance occurs when raters correlate with each other; and as discussed in the section on range restriction below, the size of a correlation coefficient is dependent in part upon the variability of the values that are being correlated. Therefore, if range restriction exists, there will be not only restricted inter-item (rater) variance but also restricted total score (subject) variance. For example, if each subject receives exactly the same score, there can be no real difference (i.e., variance) between the average scores of each subject (i.e., maximum range restriction), and the alpha coefficient will be zero. Alternatively, if each candidate receives a different score and the raters are in perfect agreement (absolutely), or in perfect consistency (relatively), the alpha coefficient will be a perfect 1.0.

As a practical example, the data in Table 3-5 appear subject to range restriction, as discussed in the next section, yet the alpha coefficient is not inflated as suggested by Fleenor et al (1996). On the contrary, as demonstrated above, Cronbach's alpha yielded the same reliability coefficient (.678) as the intraclass method (cp. Table 3-9). Another example of range restriction can be found in Tinsley and Weiss' (1975) Case 3 data, which they describe as having high interrater agreement and low interrater reliability (pp. 359-360). Analyzing this data with SPSS® Base 10.0 (1999) resulted in identical coefficients of -.28 for both intraclass *R* and alpha.

Notably, Novick and Lewis (1967) have proven that alpha is equal to or less than the true reliability of a test where the items are parallel (cited in Carmines & Zeller, 1979, p. 45). In other words, where items are not equivalent to each other, alpha will tend to

underestimate reliability and so is considered a conservative estimate of reliability. According to Green, Salkind and Akey (1997), alpha will only be overestimated if the assumption of unrelated errors is violated (pp. 359-360).[34] For example, if items on a test were somehow linked, or raters were not independent, then the alpha coefficient could be inflated.

So far, Cronbach's alpha has been compared to the intraclass method. But in the context of this study Cronbach's alpha is also noteworthy because of its relationship with the correlational approach. Under the assumption that all item variances are equal, true alpha simplifies to standardized item alpha, which is based on the average inter-item correlation (Carmines & Zeller, 1979, p. 44; SPSS® Base 10.0, 1999, p. 362):

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}},$$

which will be recognized as the Spearman-Brown correction formula introduced in the section on history.[35]

In summary, Cronbach's alpha yields the same reliability coefficient as intraclass *R* ($ICC_{3,k}$) and simplifies to the average inter-item correlation. Being a parametric index of reliability, alpha is subject to the same limitations in terms of range restriction as both the intraclass method and the correlational approach. As noted by Mehrens and Lehmann (1978), alpha is only useful when subject scores on each item take on a range of values (p. 99).

---

[34] Assumptions for reliability tests are discussed later in this chapter.

[35] Notably, by calculating total inter-item variance and a covariance-variance matrix, the results can be adapted to the Spearman-Brown correction formula to calculate true alpha (rf. SPSS® Base 10.0, p. 362).

Non-parametric Tests

To complete the discussion on reliability it is important to introduce non-parametric tests, albeit only briefly as they play a small role in assessment centers. Where rating scales produce nominal (categorical) data or strictly ordinal (ranked) data, interrater reliability assessment requires tests of agreement or concordance respectively. These non-parametric tests are described as follows.

Agreement

Logically, nominal or categorical data must be measured by interrater agreement, but such indices are relatively new additions to reliability theory (Tinsley & Weiss, 1975, pp. 361, 366). Raters either agree or disagree in their assignments of targets to qualitative categories, and so one of the first agreement measures proposed was the percentage index. Here, the number of agreements (A) and disagreements (D) are summed for each rater and the percentage agreement (P) is calculated according to the following formula: $P = [A/(A + D)] \times 100$ (Martin & Bateson, 1993, p. 120; Rosenthal & Rosnow, 1991, p. 54). Although this has an initial conceptual appeal, it is misleading because it fails to distinguish between accuracy and variability (Rosenthal & Rosnow, 1991, p. 54), although examining data in a matrix may ameliorate the problem somewhat. Additionally, percentage or proportion agreement fails to account for chance agreement, and necessarily treats agreement as absolute—a dichotomous all or none concept (Kozlowski & Hattrup, 1992, p. 163; Tinsley & Weiss, 1975, p. 366).

Cohen's (1960) *kappa* statistic (sometimes denoted as K) and Tinsley and Weiss' (1975) T-index have been described as second generation efforts to refine percentage agreement indices as they account for chance agreement (Kozlowski & Hattrup, 1992, p.

163). Cohen's *kappa* is the index most widely recommended for measuring interrater agreement on nominal or categorical data (Berry & Mielke, 1988; Cramer, 1997, p. 333; Kaplan & Saccuzzo, 1982, p. 111; Shrout & Fleiss, 1979, p. 427; Tinsley & Weiss, 1975, p. 374). Berry and Mielke also argue that *kappa* is inherently multivariate in nature and is also appropriate for interval data. Although originally designed for two raters, Berry and Mielke (1990) extended the *kappa* statistic for use with more than two raters, and Fleiss (1971) extended the *kappa* statistic for measuring interrater agreement when subjects are rated by different sets of raters (Cramer, p. 333; Tinsley & Weiss, pp. 371-372). However, a specialized computer program written by Berry and Mielke (1990) is required to calculate this expanded index, and so it is not commonly used by practitioners (see Fleenor et al., 1996, p. 379).

Notably, counting the number of agreements and disagreements of raters requires agreement or disagreement to be absolute, and treats the data as categorical. However, when the data are non-categorical, agreement does not have to be absolute (Fleenor et al., 1996, p. 369; rf. also Lawlis & Lu, 1972). According to Tinsley and Weiss (1975), their T-index, an extension of Lawlis and Lu's method and patterned after Cohen's (1960) *kappa*, is appropriate for non-categorical data and allows the researcher to decide upon what constitutes agreement above the chance level (p. 367). For example, agreement may be assumed if ratings differ by no more than 1 point.

The advantage of this type of index lies in its flexibility to measure agreement; the disadvantage is the decision on what constitutes agreement is somewhat arbitrary and subject to manipulation in order to obtain a high index of agreement. For example, on a nine point rating scale (e.g., 1, 1.5, 2, 2.5 ... 5), if agreement were defined as consensus

within one point, three raters could give assessments of 2.5, 3, and 3.5 and be considered to be in agreement because they were within one point of each other. As a result, in addition to being a function of the consensus between raters, this agreement index is, at least in part, a function of the definition of what constitutes agreement.

Notwithstanding the advantages of agreement indices, Kozlowski and Hattrup (1992) still describe them as deficient because they use a purely random model for chance response by raters (p. 163). However, this criticism also applies to classical reliability theory in general, where error is assumed to be random.

Concordance

A special form of interrater agreement may be defined as concordance, which refers to non-parametric tests of correlation of data at least at the ordinal level. Such measures are used to quantify association solely on the basis of absolute ranking or ordering where the size of intervals between points of measurement is irrelevant. For example, the earliest and probably the most widely used rank correlation is that developed by Spearman, sometimes known as "rho" and denoted as r_s (Siegel, 1956, p. 202). In practice, usually the results of Spearman's rho[36] are similar to Pearson's r, and it is sometimes used to assess interrater reliability (Berry v. Omaha, 1975; Greenwood & McNamara, 1967, p. 103; see also Miller & Whitehead, 1996, p. 329).

While Spearman's rho is concerned with the measure of association between two sets of rankings of N objects or subjects, Kendall's (1948) coefficient of concordance W measures the relation among several (k) rankings of N object or subjects.[37] As a result, W

[36] Computational examples may be found in Miller and Whitehead (1996, pp. 329-332) and Siegel (1956, pp. 202-213).
[37] For a computational example, see Siegel, 1956, pp. 229-238.

gives the average Spearman rank correlation between each pair of raters (yielding the average reliability of one judge) which makes Kendall's *W* especially useful in the context of interrater reliability (Guilford & Fruchter, 1978, p. 271; Siegel, 1956, p. 229; Tinsley & Weiss, 1975, p. 366).[38] Siegel describes this measure as "a solution to the problem of ascertaining the over-all agreement among *k* set of rankings" (p. 229), and Guilford and Fruchter (1978) describe it as analogous to intraclass *r* (i.e., $ICC_{3,1}$) (p. 271).

A coefficient of concordance, then, is a measure of association based on rank. As a result, if the purpose of an assessment center is solely to rank the candidates of a single group, then non-parametric procedures for rank correlation are appropriate. For example, a company may wish to promote the top ranking candidate(s) based on performance, or have the top ranking candidates proceed to the next stage of promotion. In such a situation, absolute agreement, relative ranking, and scores that are intended to represent a universe of all candidates are unimportant.

Special Considerations

This section will discuss three important considerations for parametric tests of reliability: underlying assumptions, range restriction, and correction for reliability attenuation.

[38] As an index of interrater reliability, Kendall's non-parametric test *W* is conceptually similar to the parametric correlational approach *r* (i.e., the average correlation between raters is calculated). The difference between the two approaches is that *W* ignores relative differences between scores (i.e., dispersion of rankings). In other words, it is limited to rank order. For a similar statistic, see also Guilford (1954), cited in Dick and Hagerty (1971).

Assumptions for Parametric Tests[39]

As previously discussed in the section on parametric tests, the assumptions for numeric and non-numeric data determine the mathematical operations that can be performed on them. Building on this discussion, this section will briefly introduce general assumptions for correlations and analysis of variance (ANOVA), followed by an outline of general assumptions for the interrater reliability procedures used in this study.

Assumptions for correlations include a linear relationship between variables, interval level data, and a joint normal distribution without outliers (Kachigan, 1991, p. 128; Koosis, 1997, p.196; Martin & Bateson, 1993, pp. 118; 141-143; Miller & Whitehead; 1996, p. 329; Vito & Latessa, 1989; Wildt & Ahtola, 1978, p. 89). Additional assumptions are that the variables are random (they covary naturally) and the scores on one variable are independent from those of the other (i.e., independence of errors) (Erickson & Nosanchuk, 1977, pp. 241, 340; Green et al., 1997, p. 282; Rosenthal & Rosnow, 1991, pp. 315, 326). A normally distributed population without outliers is important because outliers (extreme values) can spuriously inflate or distort a correlation coefficient.

Assumptions for ANOVA include homogenous variances of the sample groups, normal distributions, dependent variables at the interval level, and random samples to ensure that errors are independent (Guilford & Fruchter, 1978, pp. 283-284; Miller & Whitehead, 1996, p. 295; Rosenthal & Rosnow, 1991, pp. 315, 435). For repeated-

[39] Because non-parametric tests or statistics make few assumptions, especially concerning normal distributions, and because non-parametric tests were not a major focus of this paper, they are not discussed here. For a detailed discussion of assumptions associated with parametric tests, refer to Siegel, 1956. Although this text is over 40 years old, it is a classic. Another more recent text that offers good discussions on non-parametric tests is that of Miller and Whitehead (1996).

measures ANOVA, there are additional assumptions related to the intercorrelations among the various levels of each factor, which are assumed to be met to the degree that correlation coefficients are homogenous (Rosenthal & Rosnow, p. 435). Note, however, that for one-way repeated measures used for estimating reliability, there are no replications on the factor of interest (i.e., raters).

According to the Central Limit Theorem, a sampling distribution will tend to be normal in shape as the sample size becomes large. Specifically, experience as shown that the assumption of normality can be met when sample size exceeds thirty (Kachigan, 1991, p. 89; Miller & Whitehead, 1996, p. 295). In other words, if the sample is random and the size exceeds thirty, the researcher may conclude that the assumptions regarding normal distribution have been met; but if the sample size is less than 30, these assumptions must be more carefully examined. Accordingly, for ANOVA sample sizes of at least 30 are recommended for each group or cell (Koosis, 1997, pp. 66, 138-139; Miller & Whitehead, p. 264).

In practice, however, a sample size of 30 for each group or cell is often not possible, especially for complex designs that include a number of factors with various levels on each factor, so researchers will usually use between 10 and 15 subjects for each cell and examine the data regarding the assumptions (rf. Bruning & Kintz, 1968, Part 2, "Analysis of Variance").[40] According to Guilford and Fruchter (1978), ANOVA is insensitive to minor violations of normality and marked variances in homogeneity when there are the same number of subjects in each set (p. 284; see also Miller & Whitehead,

[40] Some statistical tests exist to test assumptions associated to various parametric techniques (see Miller & Whitehead, 1996), although often they are not satisfactory (Guilford & Fruchter, 1978, p. 284).

1996, p. 284). And according to Rosenthal and Rosnow (1991), ANOVA is described as robust, even in the face of "some fairly serious violations" of normality and homogeneity of variance (p. 326).

For most tests of interrater reliability, however, general parametric assumptions (i.e., normal distribution and homogeneity of variances) are not applicable because the tests are not techniques for statistical inference, where parameters from a sample are assessed as possibly being representative of a population.[41] Rather, tests of reliability are objective techniques for assessing, from a set of observations, the property of reliability of a particular measurement procedure. The foundation of such tests, then, is more mathematical than statistical. For parametric tests of reliability, though, there are four general underlying assumptions that should not be ignored (Green et al., 1997, pp. 359-360; Kaplan & Saccuzzo, 1982, p. 98; Li et al., 1996, pp. 98-99). First, all items are assumed to be parallel measures; i.e., equivalent to each other, having the same weight, value *r*, and variance. Second, all items are assumed to be measuring the same dimension; i.e., all items comprise a single subtest or test on a unidimensional trait. Third, all errors are assumed to be random and unrelated; i.e., ratings or scores are unbiased and independent. And last, an item score is assumed to be the sum of its own true score and error score. The first three assumptions will likely only be met approximately, while the last assumption, as previously discussed, is strictly theoretical and so the degree to which it is met cannot be known.

The purpose of this discussion was to assist in the interpretation of interrater reliability statistics, not to suggest their futility because various assumptions cannot be

[41] For example, testing for significant differences between groups.

completely met. Despite their limitations, reliability statistics provide important procedures (which admittedly cannot be fitted perfectly to operational circumstances) for assessing the decisions made by raters, which is no small matter when considering that raters make judgments that affect the future of people.

Range Restriction

A correlation coefficient cannot be calculated without variability in values and so the size of *r* is in part dependent upon the range of scores. Self-correlations in reliability studies are similarly dependent upon the range of scores in the population sample being assessed (Guilford & Fruchter, 1978, p. 324). Because in practice the reliability coefficient is a function of both real (true) differences between individuals and error differences, if no real differences between individuals exist the coefficient will be zero because differences between individuals would be due to random error only, which of course averages to zero (Fleenor et al., 1996, p. 368; Rosenthal & Rosnow, 1991, p. 50; Traub, 1994, pp. 30, 34; 110).

Causes of Range Restriction

Parametric indices of reliability such as Pearson's *r*, intraclass, and Cronbach's alpha are sensitive to range restriction (Tinsley & Weiss, 1975, pp. 362, 364; Fleenor et al., 1996, p. 370, citing James et al., 1984). Consequently, when used to measure individuals who are homogenous on the variable(s) of interest, such as the best candidates in an assessment center, "errors of measurement are most troublesome" because the errors conceal real differences between subjects (Cronbach, 1984, p. 172). If such is the case, the reliability coefficient for a homogenous group of subjects will be artificially lower than if a broad group of subjects were tested because the true score variance is in

part dependent upon the range of scores (Cronbach, pp. 172-173; Kozlowski & Hattrup, 1992, p. 163; Whitehurst, 1985, p. 568). In other words, a low reliability index may be an artifact of range restriction (see Guilford and Fruchter, 1978, p. 431, for a mathematical proof).[42]

Table 3-5 above provides an example of a restricted range of scores (34.7 to 41.375), where the distribution of scores is not evenly distributed across the rating scale (15 to 55) (rf. Table 3-1). Although this range restriction may be partly explained by central tendency (discussed later), where raters tend to assess most candidates along the mid-point of the scale, it is likely also explained by candidate homogeneity, where police recruiters because of expense only forward the best candidates to the assessment center. Theoretically, if the maximum range were used and scores were evenly distributed across categories, the range in scores would be 40.0 compared to 6.675 and the standard deviation would be 12.5 compared to 1.245, which would provide considerable more variability in scores and increase the reliability coefficient.

Although, range restriction may be in part a function of group homogeneity, it may also be a function of systematic rating errors such as leniency, central tendency, and halo effect (Borg & Gall, 1983, pp. 482-483; Kulis, 1987, p. 131; Tinsley & Weiss, p. 368). Leniency refers to raters who tend to rate all subjects the same, generally at the higher end of the scale, despite obvious differences between them. Central tendency refers to raters who tend to rate all subjects towards the middle of the scale and thereby

[42] Paradoxically, even though reliability may be low when measured in terms of consistency, it may be high when measured in terms of agreement (Fleenor et al., 1996, pp. 368, 372-373, 376; Tinsley & Weiss, 1975, p. 360). For example, Tinsley and Weiss demonstrated that one could have high interrater agreement and high interrater consistency, low agreement and high consistency, and high agreement and low consistency.

avoid being put in a position where ratings must be defended. Halo effect or reverse halo effect refers to raters who form an early opinion (or bias) on a subject (or class of subjects), this opinion unjustifiably influencing all subsequent ratings on that subject or class of subjects (e.g., sex discrimination). If these rating errors occur repeatedly, systematic error in addition to random error may be introduced into the measurement process.

Correction of Range Restriction

If systematic error is not a factor, range restriction in one population may be corrected by comparing data from another similar population where the unrestricted variance is known. For example, the following formula by Guilford and Fruchter (1978) will estimate the corrected reliability:

$$r_{nn} = 1 - \frac{S_o^2 (1 - r_{oo})}{S_n^2} ,$$

where r_{nn} is the unknown corrected reliability, S_o equals the standard deviation of the restricted distribution for which the reliability coefficient (r_{oo}) is known, and S_n equals the standard deviation in a more variable group in which the reliability is not known (pp. 325-327, 431-432). The problem, though, is that the variance of unrestricted ratings in a similar population is generally not known in assessment centers (Fleenor et al., 1996, p. 377), although if careful records were kept over an extended period under standardized conditions, it could be estimated.[43]

[43] It is assumed that the variable corrected for range restriction is linearly related to any variable to which it is correlated (Borg & Gall, 1983, p. 594).

Finn's *r*

Occasionally, when correction of range restriction is not possible, and the data are at least ordinal, Finn's *r* (1970) is recommended because it is insensitive to range restriction (Tinsley & Weiss, 1975, pp. 361-362, 374; Whitehurst, 1985, pp. 568-569). The formula is:

$$r = 1 - \frac{\text{Observed variance}}{\text{Chance (expected) variance}}, \text{ or}$$

$$r = \frac{\text{Chance (expected) variance} - \text{Observed variance}}{\text{Chance (expected) variance}},$$

where observed variance is the within subjects mean square obtained from one-way ANOVA. Chance variance (s_c^2) is the variance expected if the ratings were randomly assigned, and is calculated as follows:

$$S_c^2 = \frac{k^2 - 1}{12},$$

where *k* equals the number of points on the rating scale (Tinsley & Weiss, 1975, p. 361).

Before Finn's *r* is calculated, Tinsley and Weiss (1975) recommend that the following modified chi-square test (χ^2) be applied to determine if observed variance is smaller than chance variance by chance alone (p. 362):

$$\chi^2 = \frac{N(K\text{-}1)S_o^2}{S_c^2},$$

where *N* equals the number of subjects, *K* equals the number of raters, S_o^2 equals the observed within subjects variance (mean square from one-way ANOVA), S_c^2 equals the expected chance variance, and *N(K-1)* equals the degrees of freedom for a one-tailed chi-

square test. For this formula, the null hypothesis that observed variance is equal to chance variation is rejected if one obtains a chi-square value less than the critical value for a test at $p \leq .01$, indicating that observed variance (S_o^2) is significantly less than chance variance (S_c^2) (Tinsley & Weiss, 1975, p. 362). According to Tinsley & Weiss, Finn's r should be calculated only when the null hypothesis is rejected (p. 362).

Because it is suspected that the scores in class 245a (rf. Table 3-2) of the Police Academy assessment center are restricted in range, thereby artificially lowering the reliability coefficient, it may be useful to calculate Finn's r. First, the expected chance variance (S_c^2) must be calculated (k equals the number of points on the rating scale, which as Appendix I indicates is 13):

$$S_c^2 = \frac{13^2 - 1}{12} = \frac{169 - 1}{12} = 14.$$

Second, from the data in Table 3-4, the modified chi-square χ^2 is calculated as follows:

$$\chi^2 = \frac{6(3-1)3.78}{14} = 3.24.$$

The critical chi-square value (df = 12, $p \leq .01$) is 26.217 (rf. Rosenthal & Rosnow, 1991, p. 596). The obtained chi-square value is less than the critical value ($3.24 < 26.217$), indicating that observed variance (S_o^2) is significantly less than chance variance (S_c^2). Because the null hypothesis that observed variance was equal to chance variance can be rejected, the next step is to calculate Finn's r:

$$r = \frac{14 - 3.78}{14} = 0.73.$$

Keeping in mind that Finn's *r* refers to "the average of the reliabilities of the individual judges" (Tinsley & Weiss, 1975, p. 364), Finn's *r* must be compared with corresponding *r*'s from either the correlation approach ($\bar{r} = .44$) or the intraclass approach ($\bar{r} = .41$), both of which are sensitive to range restriction. Finn's *r*, then, indicates that the reliability for class 245a is much higher than that calculated by way of correlation or intraclass techniques, which may have been attenuated due to range restriction.

However, Tinsley and Weiss (1975) state that Finn's *r* is appropriate "only when the within-subjects variance in the ratings is so severely restricted that the intraclass correlation is inappropriate" (p. 374), which is not the case for the data here. Moreover, because Finn's *r* is a function of the number of categories on the rating scale, a rating scale with numerous categories where the extremes are not being used will artificially inflate the reliability coefficient (Tinsley & Weiss, p. 362). This appears to be true for the Police Academy assessment center, where an examination of scores over the last 22 years indicates that the extreme high points of the scale are not used. Rather, the rating scale is more accurately defined in terms of 11 points (i.e., from converted scores of 15 to 48 (rf. Table 3-1)), resulting in a Finn coefficient of .62, which is more consistent with those determined by correlation and intraclass.

Correcting Correlations for Unreliability

Correcting correlations for unreliability, otherwise known as correction for attenuation, is an important concept to apply when interpreting a correlation between two variables. When the scores of two measurement scales are correlated, the obtained

coefficient (usually Pearson's *r*) will provide an estimate of the "true" correlation only to the extent that the measurement scales are reliable (Borg & Gall, 1983, p. 593). For example, if candidate sex were correlated with overall scores at the Police Academy assessment center, but the measurement of overall scores was unreliable, the obtained coefficient would underestimate the true correlation between sex and overall scores. However, by statistically correcting for attenuation it is theoretically possible to estimate the true correlation between sex and overall scores as if measurement reliability were perfect. Because the appropriateness of correcting for attenuation for measuring test validity has been reviewed extensively in educational, psychological, and personnel selection literature, validity will be used as a convenient bridge to connect the concept of correcting for attenuation with the measurement of discrimination.

Returning to the theory of the "true score," it was previously noted that reliability may be defined in terms of the correlation between strictly (theoretical) parallel measures. Of particular interest to this discussion is that mathematicians have also shown that "the correlation between the true and observed scores is equal to the square root of the reliability," which in turn is equal to the square root of the correlation between parallel measures (Carmines & Zeller, 1979, p. 34; see also SPSS® Base 10.0, p. 365). Moreover, mathematicians have shown that the correlation between a parallel measure and some other measure cannot exceed the square root of the parallel measure's reliability (Carmines & Zeller, p. 34, citing Lord & Novick, 1968):

$$r_{xy} \leq r_{tx} = \sqrt{r_x} = \sqrt{r_{xx1}}.$$

Therefore, the square root of the reliability of a measure defines the upper limit of its correlation with another measure. In other words, the correlation between one variable and another cannot exceed the square root of the reliability of the first variable. For example, if the reliability of one variable were .6, then the correlation of that variable with another cannot exceed .775. As this logic necessarily applies to both variables, the correlation coefficient cannot exceed the square root of the product of the reliability coefficients of the variables (Kaplan & Saccuzzo, 1982, p. 90). So, for example, if the corrected or effective reliability (R) of one variable were .6 and the other .5, then their correlation coefficient (r_{xy}) cannot exceed .55 as shown below:

$$r_{xy} \leq \sqrt{.6 \times .5} = \sqrt{.3} = .55.$$

Carmines and Zeller (1979) argue that this logic demonstrates that reliability and criterion-related validity are closely related (p. 34). For example, because the square root of the reliability of a test (the predictor) is the theoretical maximum correlation of that test with another variable (the criterion), it also represents the theoretical maximum of the validity coefficient (see also Cronbach, 1984, p. 176). Therefore, it can also be argued that the degree of validity is in part a function of the degree of reliability (Kachigan, 1986, p. 219), which was noted in Chapter 1.

If reliability and validity are closely related because evidence of validity is often provided by correlating one variable with another, it can also be argued that reliability and discrimination are closely related because evidence of discrimination can be provided by showing the relationship of one variable with another. For example, if overall scores at an assessment center were significantly correlated to candidate sex (e.g., high scores

were consistently associated with male candidates), in the absence of a bona fide occupational requirement it may constitute prima facie evidence of discrimination.[44] However, the evidence would be limited by the fact that the reliabilities of the measures define the upper limit of the correlation between overall scores and candidate sex. Therefore, as with validity, it can also be argued that the degree of discrimination as measured by *r* is in part a function of the degree of reliability, if an association between two appropriate variables is established. As a result, correcting for attenuation theoretically estimates the true correlation between scores and sex, providing a more accurate assessment of discrimination.

Given that measurement error is assumed random, correcting for attenuation can be estimated by means of the formula shown below (Carmines & Zeller, 1979, p. 48):

$$r_{corr} = r_{xy} / \sqrt{r_{xx1} r_{yy1}},$$

where r_{corr} is the corrected correlation, r_{xy} is the observed (uncorrected) correlation, r_{xx1} is the reliability of variable X and r_{yy1} is the reliability of variable Y. For example, if the observed correlation between variables X (e.g., candidate sex) and Y (overall scores) were .2, and if the reliabilities of variables X and Y were 1.0 and .3 respectively, then the corrected correlation would be as follows:

[44] The concept of discrimination will be addressed in detail in Chapter 4. Statistically, in this study the concern regarding discrimination is with main effects, and residual error is assumed to be random (consistent with classical test theory). However, this raises an interesting point, where systematic bias, which is statistically defined, must be interpreted philosophically. In the final analysis, then, the issue is moral, which society formalizes through law, and as a result the statistical analyses in this study are meaningless without a clearly defined concept of discrimination.

$$r_{corr} = .2 / \sqrt{1.0 \times .3} = .2/.548. = .37.$$

Therefore, if the reliability of Y were a perfect 1.0 (i.e., without random error), rather than the observed correlation of .2 the true correlation between overall scores and candidate sex would be .37.

By carefully examining the correction formula above and Table 3-6 below, which provides an example of correcting a correlation of .2 with varying reliabilities, the relationship between the corrected correlation, the observed correlation, and the reliabilities becomes clear.

Table 3-6

Corrections for obtained correlation of .2 for varying reliabilities

	Reliabilities (*R*)									
R	**.1**	**.2**	**.3**	**.4**	**.5**	**.6**	**.7**	**.8**	**.9**	**1.0**
.1	.99	.99	.99	.99	.89	.82	.76	.71	.67	.63
.2		.99	.82	.71	.63	.58	.54	.50	.47	.45
.3			.67	.58	.52	.47	.44	.41	.39	.37
.4				.50	.48	.41	.38	.35	.33	.32
.5					.40	.37	.34	.32	.30	.28
.6						.33	.31	.29	.27	.26
.7							.29	.27	.25	.24
.8								.25	.24	.22
.9									.22	.21
1.0										.20

Note: reliabilities are in bold, and the corresponding corrected correlations are found where the rows and columns of interest intersect.

That is, as the reliabilities become higher the difference between the corrected coefficient and the observed coefficient decreases, which is illustrated by following the diagonal between .1 and 1.0 in Table 3-6 (where all corrections are for an obtained correlation of .2). Also illustrated in this table is the fact that when both reliabilities are .8 or higher, there is very little correction for attenuated correlations caused by random

measurement error. Finally, because the formula provides an estimate, when reliabilities are extremely low, an estimate of a corrected correlation of 1.0 and above can be obtained, which convention reduces to .99 (Borg & Gall, 1983, pp. 593-594).

In conclusion, in psychometric-related literature correcting for attenuation is most often discussed in the context of validity; however, it also has important implications in the context of discrimination, which may be measured in terms of a correlation coefficient. For example, discrimination may be discounted if an insignificant correlation (e.g., .2) were obtained between overall scores and candidate sex. However, by examining the reliabilities and correcting for attenuation, what was thought to be insignificant may in fact be cause for concern should reliabilities be improved. This is an important consideration, because the probative value of statistical evidence in an analysis of systemic discrimination can be expected to increase in Canadian courts and tribunals (Brook, 1990, p. 132; Vizkelety, 1987, pp. 168, 173).

Acceptable Levels of Interrater Reliability

An important question for an operational assessment center is, "How reliable must behavioral ratings be before they are acceptable?" The answer to this question depends upon the nature of the variables being measured (e.g., the difficulty associated with the assessment), the heterogeneity of the subjects (remember range restriction), the importance of the behavioral dimensions and/or the overall assessments, and the purpose of the test. With resect to minimum standards for reliability, a review of assessment center literature is useful for identifying generally accepted benchmarks.

Theoretically and operationally, reliability is defined generally in terms of a correlation coefficient, and so as a "rule of thumb" the following interpretation of *r* (apparently suggested by Guilford) is useful for interpreting the value of reliability coefficients (Martin & Bateson, 1993, pp. 143-144; see also Miller & Whitehead, 1996, p. 322):

Table 3-7

Interpreting the strength of *r*

r	Interpretation of relationship
0.0 - < 0.2	slight—almost negligible
0.2 - < 0.4	low—definite but small
0.4 - < 0.7	moderate—substantial
0.7 - < 0.9	high—marked
0.9 - < 1.0	very high—very dependable

Rosenthal and Rosnow (1991, pp. 50-51), citing Parker, Hanson, and Hunsley (1988), reported the reliability coefficients of major psychological tests used in a clinical setting. For the Wechsler Adult Intelligence Scale (WAIS), the Minnesota Multiphasic Personality Inventory (MMPI), and the Rorschach inkblot test, the internal consistency coefficients were .87, .84, and .86 respectively. According to Rosenthal and Rosnow, these reliability coefficients "are all quite respectable by psychometric standards" (p. 51). Another well known test is the Wonderlic Personnel Test (WPT), a short test of cognitive ability or general intelligence, which reports reliability coefficients ranging between .77 and .94 (Wonderlic Personnel Test, 1992, p. 47; see also McKelvie, 1989).[45]

[45] This test has recently been used by municipal police departments in British Columbia to screen police applicants.

In contrast, where assessment center raters measure behavior by recording it on various scales, as opposed to the well constructed tests of personality traits and cognitive ability noted above, reliability coefficients are generally less than .80 (Traub, 1994, p. 39). Martin and Bateson (1993, p. 119) and Kulis (1987, p. 128) recommend minimum reliabilities of .70 and .60 respectively, which they suggest are generally adequate for assessment centers. Although these authors do not provide supporting rationale, their recommendations may not be too far off the mark.

For assessment centers, reliability coefficients of .80 or greater are generally regarded as high (e.g., see Shechtman, 1992, p. 384; and Pynes and Bernardin, 1989). In the classic case of Berry v. Omaha (1975), which involved a promotional assessment center for police, a reliability coefficient of .84 (Spearman's "rho") was described as "very high" (Mendenhall, 1992, p. 62; Kulis, 1987, pp. 38-39), which appears to be an accepted interpretation.

Kulis (1987) noted that reliability coefficients as high as those found for well constructed clinical tests were sometimes reported to be common in assessment centers, but argues that this is more legend than fact (p. 128). This argument is supported by a review of the literature, summarized in Table 3-8, which shows that assessment center reliability coefficients are in fact often below .60. For example, in an early major study of assessment center reliability, when the assessment center legend was growing, Hinrichs and Haanpera (1976) reported an average coefficient of only .49.

In the absence of minimum standards set by a professional organization such as the International Congress on the Assessment Center Method (Guidelines, 1989), acceptable benchmarks are established by convention. For example, in the studies noted

in Table 3-8,[46] .60 is generally considered to be a reasonable and sufficient corrected level of reliability (*R*) for assessment centers designed for employee selection or promotion (although .60 is considered somewhat low for developmental purposes).

Table 3-8

Establishing standards for overall reliability (corrected *R*) for assessment centers

	Studies											
Standards	1	2	3	4	5	6	7	8	9	10	11	12
less than adequate < .30	X			X					X			
marginal adequate ≥ .30 < .45	X			X			X					
low adequate ≥ .45 < .60	X			X			X					
adequate ≥ .60 < .75	X		X	X	X	X	X	X	X	X		
high adequate ≥ .75 < .90	X	X	X	X	X	X		X			X	X
very high adequate ≥ .90	X	X		X	X	X		X				

Note: "X" indicates that the study, noted at the top of the column, explicitly or implicitly in some way discussed what constitutes accepted reliability. This table has attempted to translate these discussions into meaningful classifications, which are noted in the far left column.

Legend:

1 Adams & Thornton (1988)
2 Berry v. Omaha (1975)
3 Greenwood & McNamara (1967)
4 Hinrichs & Haanpera (1976)
5 Howard (1974)
6 Huck (1977)
7 Hutton & Sampson (1999)
8 Kudisch et al. (1997)
9 Kulis (1987)
10 Martin & Bateson (1993)
11 Parker (1980)
12 Traub (1994)

Nevertheless, assessment center administrators should strive for reliability of .80 because at that level corrections for attenuated correlations are minimized (see Table 3-6), ensuring that observed correlations better reflect true correlations when measuring validity or discrimination.

[46] These studies also summarize the reliabilities reported by other important studies. For example, see Howard (1974) and Hinrichs and Haanpera (1976), who report the reliability coefficients for Bray's MPS.

Summary and Recommendations

As stated by Rosenthal and Rosnow (1991), "If we are to understand the functioning of a test, we must understand its reliability" (p. 47). In response, this chapter first provided the framework (i.e., classical reliability theory) in which to understand interrater reliability in the context of an assessment center. Second, this chapter demonstrated the importance of reliability to an assessment of discrimination (i.e., when measured by *r*, the level of observed discrimination is in part a function of the reliabilities of the variables of interest). Third, because reliability cannot be generalized to the assessment center method, this chapter provided assessment center administrators with a practical and uncomplicated yet suitable parametric method to estimate interrater reliability. And finally, applying three more strokes of Occam's razor, this chapter concludes by briefly summarizing the theoretical issues, suggesting a practical guide to the methodology maze, and providing an enhanced rationale for the correlational approach.

Theory

This chapter has focussed on classical measurement theory because of its dominance in measurement and because it provides a suitable framework in which to understand interrater reliability. Because interpretations of consistency and agreement account for a great deal of the confusion in reliability measurement, the debate regarding these concepts was briefly reviewed. While it is understood that agreement and consistency may be conceptually distinct, it is not accepted that they are conceptually incompatible. Consistency and agreement are admittedly different, but both find their place within reliability as both are concerned with measurement accuracy. Although the

debate continues, it is suggested here that the argument may be resolved, at least at a practical level, by accepting the inherent limitations of their respective methodologies. Different data and purposes demand different methods, but the overarching issue is still one of reliability in which the concern is measurement error.

Methodology

This chapter reviewed the most common parametric indices of interrater reliability, which are the correlational approach, the intraclass method, and Cronbach's alpha. In addition, this chapter addressed important methodology issues related to the choice of a reliability index, which depends upon the purpose of the study and the data available. Admittedly a complicated subject that causes a great deal of confusion, the suggested "methodology map," as illustrated in Figure 3-1, will hopefully clarify the search for the appropriate reliability technique.[47]

As indicated in the methodology map, data type is the most important factor to consider when selecting an interrater reliability technique. Specifically, for non-numeric data, non-parametric tests are appropriate, and for numeric data, either non-parametric or parametric tests are appropriate. For example, when the relative consistency between raters is of interest, and the data type is at least quasi-interval (ordinal data where measurement intervals are logically equal), the correlational approach, Cronbach's alpha, or the intraclass technique is appropriate.

[47] There is no complete consensus about how the various indices should be categorized (all part of the confusion in the field). For example, Tinsley and Weiss (1975) debate the fact that Lu (1971) describes his reliability index *A*, which is very similar to Finn's *r*, as an index of agreement (p. 369). The purpose of this methodology map, then, is not to resolve the debate, but to make practical suggestions.

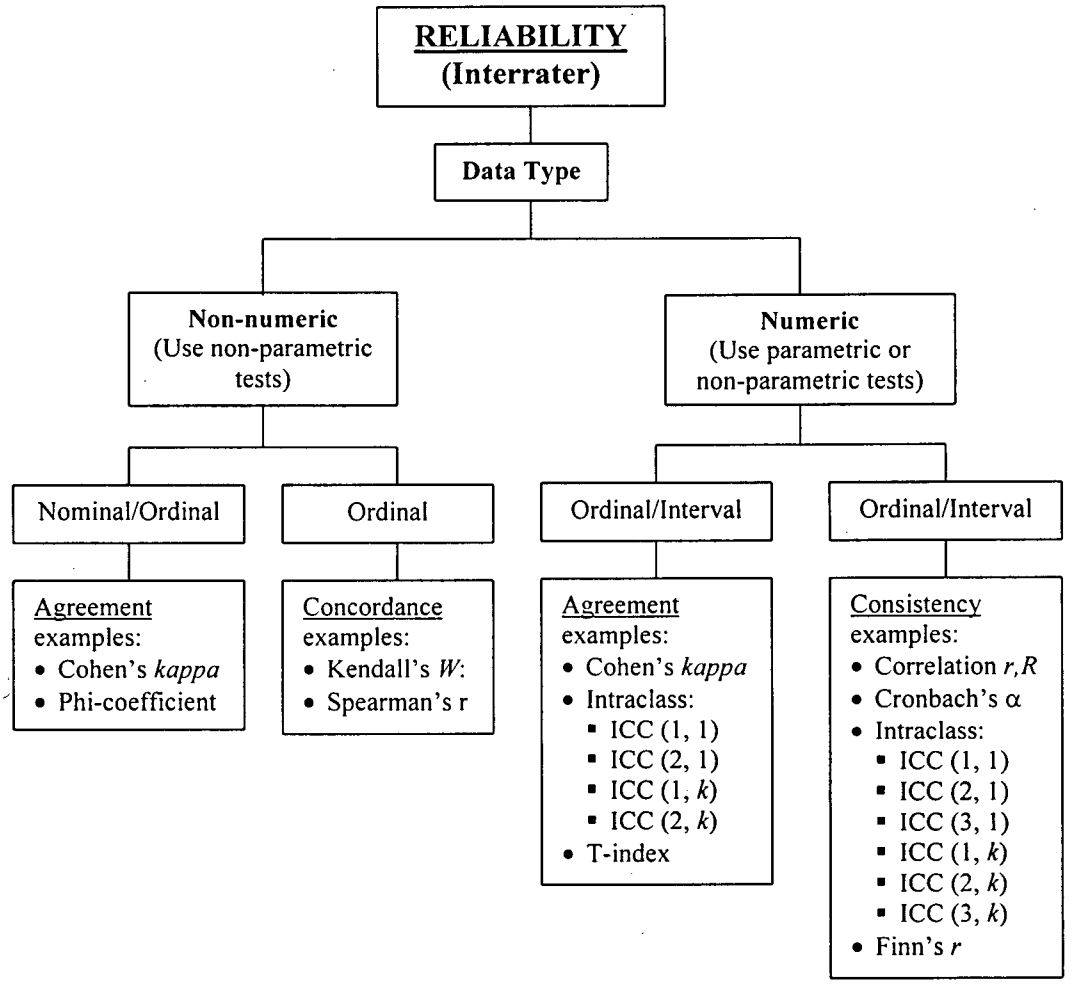


Figure 3-1. Methodology map.

Recommendations

When raters are not too numerous (e.g., less than four or five), as is the case at most assessment centers, the correlational approach is an appropriate method for estimating interrater reliability. An appropriate alternative is Cronbach's "true" alpha (based on covariance-variance),[48] where the formula is easily followed, although the

[48] Standardized alpha simplifies to the correlational approach.

calculations necessary for the summary statistics to use in the formula are somewhat onerous. The disadvantage of this method is that for those not trained in statistics it lacks the conceptual clarity of the correlational approach.

If there are a large number of raters and subjects the correlational approach can become onerous (but not complicated), and so Cronbach's alpha or the intraclass approach may be more expedient. However, in addition to the disadvantages of Cronbach's alpha, the intraclass approach has the disadvantage of six different versions and complex calculations, although in a computerized environment this is not necessarily a problem for either Cronbach's alpha or intraclass. A computer program such as SPSS® offers speed and efficiency, but its chief disadvantage (apart from expense) is that for those not trained in statistics the conceptual clarity is again lacking.

Although the correlational approach is criticized by some (e.g., Fleenor et al., 1996, p. 377-378), many researchers describe it as appropriate (Carmines & Zeller, 1979, pp. 43-48; Cramer, 1997, p. 336; Greenwood & McNamara, 1967, p. 102; Mehrens and Lehmann, 1978, p. 101; Rosenthal & Rosnow, 1991, pp. 51-54; Rosnow & Rosenthal, 1996, pp. 127-128; Traub, 1994, p. 73). Of particular importance is that this study has demonstrated that differences between reliability coefficients calculated by the correlational approach (which overestimates slightly) and the intraclass approach and Cronbach's alpha are insufficient to lead to erroneous conclusions about reliability. For example, the results of an analysis of Police Academy class 253, reported in Table 3-9 below (which replicate the results reported in Tables 3-2, 3-3 & 3-5), show that the differences between the approaches are small, similar to that found by Cronbach (1984, pp. 169-170), Fleenor et al. (1996, pp. 377-378) and Rosenthal and Rosnow (1991, pp.

55-56). So, for practical purposes, the evidence suggests that the correlational approach is an appropriate method for assessing interrater reliability.

Table 3-9

Academy class 253: Comparing reliability coefficients

Part I: Rater Scores

	Raters		
Candidates	A	B	Total
1	39.100	40.167	79.267
2	32.417	34.958	67.375
3	30.833	34.167	65.000
4	36.417	34.000	70.417

Part II: Reliability

	Coefficients	
Technique	Single $\bar{r}$	Corrected R
Correlational approach	.74	.85
Intraclass (ICC)	.72	.83
Cronbach's alpha (α)		.83

The advantages of the correlational approach, then, are that it is reasonably accurate, easily calculated, and conceptually clear to those not familiar with analysis of variance and related formulae,[49] yet it is consistent with the strict definition of reliability.[50] A person requires no training in statistics in order to understand the concept of correlation, by which one variable is seen to be associated with another, and coefficients are easily calculated by using an inexpensive hand calculator. Because of its convenience, the correlational approach can be used for every assessment center class in

[49] The main advantage of the intraclass approach is that the variance associated with each component is identified. These advantages, however, are more for researchers than practitioners.

[50] Strictly speaking, "the ICC is the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target" (Shrout & Fleiss, 1979, p. 422). Thus, the correlational approach should still have a conceptual appeal for the reliability purist because of its fit with classical reliability theory.

order to ensure that reliability meets accepted standards (i.e., at least a corrected reliability of .60).

One more advantage worth noting is that the correlational approach allows the assessment center administrator to examine the correlation coefficients between each rater (e.g., A with B, A with C, and B with C) and so assess the performance of individual raters. If an individual rater consistently does not correlate well with other raters (most likely due to systematic error, such as bias or central tendency), remedial training may be considered. And if the rater does not improve, the administrator can make an informed decision regarding the continued participation of that rater in future assessment centers.

As a parametric index the correlational approach has limitations, such as sensitivity to range restriction and outliers, but these limitations also apply to the intraclass approach and Cronbach's alpha. Another limitation is that parametric indices do not generally measure absolute agreement, which however is not an issue if the question is one of consistency or relative ordering. These limitations, though, do not preclude the use of the correlational approach, as all approaches to reliability have limitations and require cautious qualitative analysis before making interpretations (Whitehurst, 1985, p. 569).

From a quantitative perspective, the foundation has now been laid to measure interrater reliability and discrimination at the Police Academy assessment center. As stated by Shrout and Fleiss (1979), "It is important to assess the reliability of judgments made by observers in order to know the extent that measurements are measuring

anything" (p. 427). However, to meaningfully assess discrimination it is necessary to define it, both theoretically and operationally, which is the purpose of the next chapter.

CHAPTER IV: DISCRIMINATION

Research on organizational personnel selection has focussed primarily upon increasing the rate of valid selections, the aim of which is to optimize "human resources utilisation with the goal of maximising organizational productivity and effectiveness" (Singer, 1993, p. 1, 3, 29, 87). Within this utility framework, the subject matter of personnel selection consists of job analysis, criterion and predictor validity issues, and efficiency. In other words, selection research has generally been examined from a psychometric and utility perspective, and comparatively little effort has been made in systematically applying theories of justice and equality. For example, within the context of police assessment centers, "large resources continue to go into the design and implementation of programmes," while basic human rights issues are generally not addressed (Feltham, 1988, p. 142).

In the employment context, one of the most important human rights issues is that of discrimination, an understanding of which is essential in order to guide discussions on just personnel selection. With this in mind, this chapter will be divided into the following two sections: "The Idea of Justice and Equality," and "Discrimination: The Legal Concept." The first section will briefly introduce the concepts of justice and equality to set the scene for a discussion of discrimination in a legal context. The second section will provide a comprehensive analysis of discrimination within Canada's legal system, the purpose being to create a conceptual framework for assessing sex discrimination at the Police Academy assessment center.

The Idea of Justice and Equality

In order to be coherent, theories of morality, at minimum, depend upon reason and the principle of equality (Rachels, 1999, pp. 15, 18; Vlastos, 1970, p. 86). According to Wasserstrom (1970), "The principle that no person should be treated differently from any or all other persons unless there is some general and relevant reason that justifies this difference in treatment is a fundamental principle of morality, if not rationality itself" (p. 103; see also Rachels, p. 94). Therefore, to be moral one must acknowledge that the welfare of others is as important as one's own.

Justice, Equality, and Morality

As equality is fundamental to morality, indeed rationality, so it is to the concept of justice, for without equality justice loses its meaning. For example, the great struggles for justice throughout history have been about equality, such as the struggles against slavery, dictatorships, sex discrimination, etc. Singer (1993) argues that not only is justice a central concern of people, it is also basic human motive, citing Lerner's (1982) study of the justice motive in human relationships. Similarly, Watson and Barber (1988), in their historical review of democracy, found that "from the time the human race was young, people have looked to constitutions and systems of law not only for rules and regulations but for fairness and justice" (pp. 121-22). Moreover, Watson and Barber argue that since the arrival of democracy in ancient Greece justice has been conceived as a "tolerant and objective fairness," where impartiality is the cornerstone (p. 123; see also Frankena, 1963, p. 40).

Aristotle is quoted as saying in the Nicomachean Ethics, "Justice is equality" (Vlastos, 1970, p. 76), a conclusion that has been generally accepted by philosophers.

For example, according to Frankena (1963), equality and justice are one and the same principle (p. 38). Similarly, Sidgwick (1907) explained the principle of justice as follows:

> It cannot be right for A to treat B in a manner in which it would be wrong for B to treat A, merely on the ground that they are two different individuals, and without there being any difference between the natures or circumstances of the two which can be stated as a reasonable ground for difference of treatment.
> (Sidgwick, 1907, cited in Frankena, 1963, p. 15)

The classic case of injustice is when two similar individuals in similar circumstances are treated differently, which is why Sidgwick suggested his formula of justice (Frankena, 1963, p. 39). Although this commonly accepted formula is appealing in its simplicity, problems arise in its application because rarely are circumstances the same, which may then provide reasonable grounds for differential treatment. Equality, then, is not necessarily about absolute sameness; rather, it is about proportional sameness, where benefits are distributed "equitably" according to merit, work, need, and worth (i.e., distributive justice).

As a result, there is continual controversy on what is a benefit, on what benefits qualify on moral grounds, and on what is a just distribution of benefits (which are distinguished from resources). Although it is beyond the scope of this study to explore this subject, there is general agreement that to qualify on moral grounds such benefits must be related to a person's worth and well being (Vlastos, 1970, p. 86), although this concept itself lacks specificity (Whitehead, 1933, pp. 13-14). Frankena (1963) argues that the principle of justice imposes an obligation to help each person in proportion to their needs and abilities if it means that in doing so he or she would be assisted equally in the achievement of the "good life" (pp. 40-41). In a just society, then, if there are natural

human rights, such as a right to the "good life," such rights must be possessed equally by all human beings (Wasserstrom, 1970, p. 100).

Therefore, in any theory of moral obligation or duties, such as that created by a theory of human rights, justice plays an integral part (Frankena, 1963, p. 38). For example, Article One of the United Nations' Universal Declaration on Human Rights (1948) states, "All human beings are born free and equal in dignity and rights" (rf. Davies, 1988). In the legal arena, the Canadian Charter of Rights and Freedoms[1] states that every individual has the right to "equal benefit of the law without discrimination and, in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental or physical disability" (s. 15). Notably, such rights are not absolute; rather, they are, as Frankena states, "prima facie," for there are occasions when they may be overruled (p. 41). For example s. 1 of the Charter, known as the "notwithstanding clause," guarantees the rights and freedoms set out therein "subject only to such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society."

Distributive Justice and Procedural Justice

Considering the significance of justice and equality in personnel selection decisions, they ought to be a primary consideration of all personnel managers. Singer (1993), an organizational justice theorist, suggests two basic questions for personnel managers: what are just outcomes, and what are just procedures (p. 3)? These questions illustrate two conceptually distinct subsets in the discussion on justice: distributive justice (e.g., see Deutsch, 1985), introduced above, where the emphasis is on outcome or

[1] Hereafter referred to as the Charter.

benefits, and procedural justice (e.g., see Lind & Tyler, 1988), where the emphasis is on process. Despite these differences, however, distributive justice and procedural justice are inextricably related, similar to substantive law and adjective law.[2] For example, in a legal context, equality can be both substantive where discrimination is prohibited, and procedural, where all persons are entitled to an impartial decision. Rawls (1993), in his text Political Liberalism, stated:

> Both kinds of justice exemplify certain values, of the procedure and the outcome, respectively, and both kinds of values go together in the sense that the justice of a procedure always depends on the justice of its likely outcome. Thus, procedural justice and distributive justice are connected and not separate. (p. 421)

As in distributive justice, there is general agreement in the definition of procedural justice, whether from an organizational perspective or a legal perspective. For example, from an organizational perspective, common indicators of procedural justice include consistent standards, safeguards against bias, appeal provisions, impartiality, openness, and use of relevant information (Leventhal, Karuza & Fry, 1980; Rawls, 1993; Singer, 1993). Using Thibaut and Walker's (1978) theory of procedure, Singer classified his procedural indicators along two dimensions: process criteria (e.g., honest communication, choice processes, and use of relevant information), and decision factors (e.g., job relevance and bias avoidance) (pp. 51, 87-88). This conceptualization of procedural justice in terms of process criteria and decision factors is substantially similar to that found in law. For example, from a Canadian legal perspective, "natural justice,"

[2] Substantive law represents society's attempt to codify (operationalize) distributive justice, just as adjective law represents society's attempt to codify (operationalize) procedural justice. "Substantive law," or "positive law," creates and defines rights, duties, and prohibited acts, and is distinguished from "adjective law" (otherwise described as "procedural law," "due process," "procedural fairness," or "natural justice"), by which substantive law is made effective or is determined (Yogis, 1983).

or "procedural fairness," consists of two basic components: the right to be heard, and the right to an impartial decision (Jones & de Villars, 1994, pp. 180-181).

It is noteworthy, from a legal perspective, that natural justice historically referred to the exclusive requirement of judicial and quasi-judicial bodies[3] to be procedurally fair in reaching a decision, and therefore did not apply to legislative or executive bodies such as administrative tribunals (Jones & de Villars, 1994, p. 179). Recently, however, this "duty to be fair," as either an extension of the principle of natural justice or simply a restatement, has evolved in Canada where the court has asserted its power to review governmental executive decisions, excluding legislative decisions, that are not of a judicial nature (Jones & de Villars, pp. 180; 193-208). The distinction, then, between natural justice and the duty to be fair was not conceptual, but one of jurisdictional application, as stated by Mr. Justice Dickson in Martineau v. Matsqui Institution Disciplinary Board (1980):

> In general, courts ought not to seek to distinguish between the two concepts, for the drawing of a distinction between a duty to act fairly, and a duty to act in accordance with the rules of natural justice, yields an unwieldy conceptual framework. (p. 623)

Summary

Defined generally, the concept of justice, inextricably interrelated with that of equality, is moral in nature. More specifically, justice may be defined in terms of distributive justice (e.g., equitable treatment) and procedural justice (e.g., impartial treatment). In Canadian law, the concepts of natural justice and fairness overlap. Characterized by impartiality, these concepts generally refer to rules or procedures that

[3] Such bodies have authority to interpret and apply the law, including the authority to extinguish or modify private rights or interests in favour of another (Jones & de Villars, 1994, p. 186).

govern how substantive decisions are reached. Therefore, in a legal context the concept of justice, described as "fundamental justice" in the Charter (s. 7), has a procedural orientation and is concerned with principles that ensure people, individually and collectively, are treated fairly (see also s. 1; and also Tanovich, 1999, pp. 4-27). With respect to the concept of equality, in Canadian law it is defined substantively (e.g., s. 15 of the Charter), about which more will be said in the following section on discrimination.

Although conceptually distinct, distributive and procedural justice are related and not easily separated in practice. For example, whether or not the Police Academy assessment center treats men and women equally may be either a distributive issue (e.g., in terms of equitable outcomes) or a procedural issue (e.g., in terms of impartial decisions). However, whether or not the assessment center is reliable is probably more a procedural issue (e.g., in terms of consistent scores). Because an objective of this study is to determine whether, contrary to human rights legislation, the Police Academy discriminates on the ground of sex, the issue in law is substantive and therefore falls more within the general definition of distributive justice, although for practical purposes such characterizations are unnecessary.[4] Moreover, although in law the substantive issue of equality is more narrowly defined in terms of discrimination (i.e., differential treatment on prohibited grounds), it is consistent with how equality is conceived generally[5] and provides a conceptual framework to assess sex discrimination that is of practical significance in the employment context.

---

[4] Albeit a technical point, because the Police Academy is not a judicial body or an administrative tribunal, whether or not it discriminates is in law necessarily a substantive issue.

[5] Notwithstanding that in philosophy everything is controversial, where "'competent' philosophers will disagree even about fundamental matters" (Rachels, 1999, p. xii).

Discrimination: The Legal Concept

The idea of human rights dates back to early religious, philosophical, and legal theories of the natural law, which is generally believed to supercede the positive laws of sovereign states (e.g., see Martin Luther King's Letter from Birmingham Jail, 1963). In the late 18th and 19th centuries, national constitutions began to recognize human rights (e.g., the American Bill of Rights of 1789), and in 1948, after the atrocities of World War II, the United Nations General Assembly adopted the Universal Declaration of Human Rights (Weissbrodt, 1988, pp. 1-2; Davies, 1988, p. vii). In a contemporary legal context, the Universal Declaration of Human Rights is a convenient historical landmark for a discussion of human rights. While this Declaration limited the concept of human rights to discrimination on a few prohibited grounds (e.g., race, sex, language, or religion), it was an important benchmark for evaluating national policies on the issue of human rights (Davies pp. vii, 2).

Notwithstanding the limitations of legislation, the purpose of codifying human rights is "to achieve equality and eliminate discrimination" (British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union, 1999,[6] paras. 2, 81; see also Black, 1994, pp. 4, 14, 79, 156; Hurley, 1998, p. 4; and Sheppard, 1993, p. 4). For example, British Columbia's Human Rights Code (1996) states that its purpose is to promote human dignity and equal rights for all members of society, to prevent discrimination on prohibited grounds[7] (e.g., race, religion, sex, etc.), to be proactive in eliminating systemic discrimination, and to provide a

[6] Hereafter referred to as Public Service (1999). Note that at the time of writing, this case was only available in electronic format (QL Systems), so pinpoint references are by paragraph (i.e., para.).
[7] The legal concept of prohibited grounds will be addressed in the section entitled "Discrimination."

mechanism of redress for those who are victims of discrimination (s. 3). Similar legislation is found across Canada, where the federal government and all territories and provinces have enacted human rights statutes.

While the concept of discrimination is constantly debated generally, the Supreme Court of Canada has the authority to settle "fairly well" the matter legally (Andrews v. Law Society of B.C., 1989, p. 579). The problem, though, is that the legal interpretation of discrimination is complex and somewhat ambiguous, which gives rise to problems of application (e.g., see Public Service, 1999). Therefore, within this legal framework, using British Columbia's Human Rights Code as an example of provincial legislation, the overall purpose of this section will be to clarify the fundamental principles underlying discrimination as it relates to employment law in Canada, so that the issue of systemic sex discrimination may be properly understood and applied within the context of this study. Specifically, this section will outline the legislative framework that regulates human rights legislation, provide an historical analysis of the legal concept of discrimination, describe exceptions in which discrimination is justified, discuss the problem of practical application and analyze the Supreme Court's new "unified approach," and conclude with an explanation of how systemic sex discrimination may be assessed in an applied context.

Despite what may appear to be a limited discussion of discrimination (i.e., limited to personnel selection), the legal concept is essentially the same regardless of the context. Most information regarding discrimination is found in the employment context, and "there is every reason to believe that groups who experience inequality in employment face similar patterns of inequality in other areas as well" (Black, 1994, p. 4; see also

Sheppard, 1993, p. 1). Similarly, because federal, territorial and provincial human rights statutory models have the same broad purpose and a similar structure, British Columbia's legislation has general applicability.

Legislative Framework for Human Rights

Discrimination in the employment context and elsewhere is addressed by s. 15 (equality rights) of the Charter, provincial human rights legislation (e.g., British Columbia's Human Rights Code,[8] 1996), and federal human rights legislation (e.g., the Canadian Human Rights Act, 1985). Substantively, the most significant difference between s. 15 of the Charter and provincial or federal legislation is that of orientation, which is of no practical significance as both are "aimed at the same general wrong" (Public Service, 1999, para. 48). Authoritatively, the most significant difference between the Charter and provincial or federal human rights legislation is that the Charter is constitutional law, and as a result all other law must be consistent with it (see Funston & Meehan (1994, pp. 48-49) and Hogg (1985, p. 93) for a discussion on the rationale).

Specifically, s. 52(1) of the Constitution Act (1982) states that "the Constitution of Canada is the supreme law of Canada, and any law that is inconsistent with the provisions of the Constitution is, to the extent of the inconsistency, of no force or effect" (see R. v. Big M. Drug Mart Ltd. (1985) and Reference re Language Rights under S. 23 of Manitoba Act, 1870 and S. 133 of Constitution Act, 1867 (1985) for a discussion by the Supreme Court of Canada on this issue).[9] A law may be statute, regulation, or

[8] Hereafter referred to as B.C.'s Human Rights Code.

[9] As noted by Funston and Meehan (1994), the courts had the power to vitiate laws inconsistent with constitutional law prior to the Constitution Act of 1982 (p. 49), s. 52 simply a codification of previously existing authority.

common (Little Sisters Book and Art Emporium v. Canada, 1998, p. 502),[10] and the courts have not been reluctant to exercise their authority to strike down or modify any law that is inconsistent with constitutional law.

Notably, federal and provincial human rights legislation is considered to be quasi-constitutional, a status necessary to ensure its effectiveness. For example, s. 4 of the B.C.'s Human Rights Code states, "If there is a conflict between this Code and any other [provincial] enactment, this Code[11] prevails." Moreover, the Supreme Court of Canada has ruled that because the "Human Rights Act is legislation declaring public policy," private parties may not exempt themselves from its provisions (Winnipeg School Division No. 1 v. Craton, 1985, citing its decision in Ontario Human Rights Commission v. Etobicoke, 1982; rf. also Black, 1994, p. 152). Without exception the Supreme Court has affirmed this elevated status (e.g., Insurance Corporation of British Columbia v. Heerspink, 1982; and Zurich Insurance Ltd. v. Ontario (Human Rights Commission), 1992),[12] as explained by Mr. Justice Sopinka in Zurich:[13]

> In approaching the interpretation of a human rights statute, certain special principles must be respected. Human rights legislation is amongst the most pre-eminent category of legislation. It has been described as having a "special nature, not quite constitutional but certainly more than ordinary ..." (Ontario Human Rights Commission [& O'Malley] v. Simpson Sears Ltd., [1985] 2 S.C.R. 536, at 547). One of the reasons such legislation has been so described is that it is often the final refuge for the disadvantaged and the disenfranchised. (p. 339)

---

[10] The Supreme Court of Canada has held that for s. 15 of the Charter, "law" is not confined to legislation such as laws and regulations, but also includes government policies and contracts (Hurley, 1998, citing McKinney v. University of Guelph, 1990).

[11] According to Black (1994), the term "code" vis-a-vis "act" has symbolic significance, indicating superior status (p. 33).

[12] See also Black, 1994, pp. 2, 33, 153-154; Sheppard, 1993, p. 3; and Soltan, 1994, p. 4.1.01.

[13] Another important case is that of Ontario Human Rights Commission & O'Malley v. Simpson-Sears Ltd. (1985) noted in the quote. This case is sometimes referred to as Ontario Human Rights Commission v. Simpson Sears, and sometimes as O'Malley v. Simpson Sears. In this paper, this case will be referred to as O'Malley v. Simpson Sears, or simply O'Malley, except in the "References" section, where the full citation is given.

What makes provincial or federal human rights legislation "quasi" constitutional is that its jurisdiction is limited; it can be changed as easily as any provincial or federal law, and it must conform to constitutional law (e.g., the Charter). Moreover, provincial legislation is subject to the doctrine of paramountcy. Here, if provincial and federal law conflict on the same matter, federal law prevails to the extent of the conflict, similar to the concept of "supreme law" by which constitutional law prevails over all other law (Funston & Meehan, 1994, pp. 51-52, 123; Hogg, 1985, p. 354; see also Deloitte, Haskins & Sells Ltd. v Alberta (Workers' Compensation Board), 1985).

Legislative duplication by itself, however, does not necessarily lead to conflict (Funston & Meehan, 1994, pp. 51-52). For example, s. 32 of the Charter restricts its application to public government agencies and to those agencies that perform a public government function (Godbout v. Longueuil (City), 1997), while provincial and federal human rights legislation applies to both public and private agencies.[14] Although provincial and federal human rights legislation overlap with each other, including s. 15 (equality rights) of the Charter, each statute is valid.[15] However, should a conflict exist, the doctrines of paramountcy and supreme law apply.

Apart from the doctrine of paramountcy, the only real difference between federal and provincial human rights legislation is that of jurisdiction. For example, in an employment context, the federal statute (Canadian Human Rights Act, 1985) regulates companies or employers that are federal in nature (e.g., the Royal Canadian Mounted

---

[14] See Black (1994, p. 12), citing Bhadauria v. Board of Governors of Seneca (1981), and Zinn & Brethour, (1996, p. 1:1).

[15] Overlap vis-a-vis conflict can also occur within a province. For example, a grievance filed pursuant to labour law may appear to be as appropriate as a complaint filed pursuant to human rights legislation (e.g., Public Service, 1999). In such cases, a number of provincial human rights statutes have provisions where the human rights complaint can be deferred until it can be determined whether the complaint was adequately addressed (e.g., s. 25 of B.C.'s Human Rights Code) (rf. Black, 1994, pp. 94, 97-99, 146-147).

Police). Provincial legislation, such as B.C.'s Human Rights Code, regulates all other companies or employers located within the province in which the statute exists (e.g., the JIBC Police Academy).

Canadian constitutional law (which is really a framework by which law making is regulated) authorizes the federal and provincial governments to enact human rights legislation,[16] which is a classic example of administrative law. Here, the government grants authority to a distinct administrative body, separate from the judicial system, to administer the statute and exercise quasi-judicial powers by way of administrative tribunals (Jones & de Villars, 1994, pp. 3-5; Gall, 1995, p. 26; Waddams, 1997, p. 60).[17] For example, B.C.'s Human Rights Code established a "human rights commission," which receives and investigates complaints (Part 3, Complaints), and a "human rights tribunal," which adjudicates complaints referred by the commission (Part 4, Human Rights Tribunal) (Black, 1994, p. 108; Gordon, 1997, p. 3:1:01). A decision and remedy of the tribunal, filed with the Supreme Court of the province, has the "same force and effect ... as if it were a judgment of the Supreme Court" (Human Rights Code, s. 39; Court Order Enforcement Act, 1996; rf. Black, 1994, pp. 25, 120-122, 141).

As often the case in such enabling statutes, there is no defined appeal procedure; rather, decisions of administrative tribunals are appealed by way of a judicial review. For example, in British Columbia, a petition for a judicial review is filed with the Supreme Court of the province (rf. Judicial Review Procedure Act, 1996). However,

---

[16] Rf. Constitution Act, 1867, ss. 91-92.

[17] Black (1994) argues that for advancing human rights the administrative model is more effective than the judicial model because the administrative model is more flexible and its purpose is more preventative than punitive (pp. 62, 65, 108, 112, 141-142). Moreover, decisions are made on a "balance of probabilities" vis-a-vis "beyond a reasonable doubt" (rf. Ontario Human Rights Commission v. Etobicoke, 1982, p. 503).

notwithstanding the existence of judicial review legislation, ordinary courts have always exercised an "inherent" jurisdiction to supervise the legality of actions by administrative officers, tribunals, etc. (Jones & de Villars, 1994, p. 9). Technically, a judicial review is not an appeal, which is broad in scope, but a review of "whether the administrator has acted strictly within the powers which have been statutorily delegated," which is relatively narrow in scope (Jones & de Villars, p. 7). A judicial review cannot review questions of fact unless they are "patently unreasonable" (Ross v. School District No. 15, 1996), and as a result is generally limited to questions of law. Furthermore, if a finding of fact is held upon review to be patently unreasonable, it is considered to be an error in jurisdiction (British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights, 1997, citing Blanchard v. Control Data Canada Ltd., 1984).

Discrimination: An Historical Analysis of the Legal Concept

Generally, discrimination towards people can have either positive or negative connotations. In Funk & Wagnalls Standard College Dictionary (Landau, 1974), "discrimination" is defined as acting with partiality or prejudice, drawing a distinction, distinguishing, differentiating, and discernment. People discriminate between others every day, whether informally in their private lives (e.g., giving preferential treatment to family and friends over strangers), or formally in their professional lives (e.g., employers making hiring decisions based on selection procedures). Discriminating between people, then, is a necessary part of society, but due to its relational nature it is subject to moral judgments—i.e., it is either justified or unjustified.

Legal Definition

The definition of discrimination is open to many interpretations, and evolves by a slow process of rational argument and debate. Although the purpose of this chapter is to provide a framework in which to understand discrimination in the legal context, no final resolution of definitional issues will be achieved because they are political as well as intellectual. In a the human rights context, discrimination usually has a negative moral connotation—i.e., it addresses issues of unfairness and inequality in terms of behavior of one person to another, which are generally considered to be unjustified or wrong. The Canadian Law Dictionary (Vasan, 1980), defines discrimination as a "failure to treat all persons equally, generally arising out of a prejudice against a class of persons" on the basis of race, colour, sex, religion, etc. (cp. Black, 1990, Black's Law Dictionary). Despite the usefulness of such general definitions, Zinn and Brethour (1996) argue that "a meaningful definition of discrimination is as elusive as a definition of justice," noting that the definition has changed significantly over time (p. 1:2; see also Black, 1994, pp. 154-155; and Hurley, 1987, pp. 26-27).[18] Because a concise definition of discrimination is generally not found in legislation (Zinn & Brethour, p. 1:2), a discussion of what constitutes unjustified discrimination must begin with the purpose of human rights legislation (Sheppard, 1993, p. 3). In Zurich Insurance Ltd. v. Ontario (1992), Madame Justice L'Heureux-Dubé stated the following:

> The starting point for any analysis of human rights legislation is the recognition that the purpose of such legislation is the protection of fundamental individual rights. These rights are violated if stereotypical group characteristics are ascribed to individuals. (p. 358)

[18] For example, it was once thought that intent was necessary to prove discriminatory conduct, and that the remedy was equal treatment, both of which have been rejected by the courts (discussed later in this paper).

Andrews v. Law Society of B.C. (1989) was the first case decided by the Supreme Court of Canada on s. 15 (equality rights) of the Charter, and is the most cited case in human rights law (e.g., over 40 Supreme Court of Canada decisions) and in the literature (e.g., Black, 1994; Hurley, 1998; Lovett, 1997; Sheppard, 1993; Zinn & Brethour, 1996) regarding the definition of discrimination.[19] In this case Mr. Justice McIntyre explained the concept, whether in the context of human rights legislation or s. 15 of the Charter, as follows:

> Discrimination may be described as a distinction, whether intentional or not,[20] but based on grounds relating to personal characteristics of the individual or group, which has the effect of imposing burdens, obligations or disadvantages on such individual or group not imposed upon others, or which withholds or limits access to opportunities, benefits, and advantages available to other members of society. Distinctions based on personal characteristics attributed to an individual solely on the basis of association with a group will rarely escape the charge of discrimination, while those based on an individual's merits and capabilities will rarely be so classed. (pp. 174-175)

In provincial human rights law the defining question is whether discrimination is based on prohibited grounds (which are essential characteristics of discrimination) in specified areas, such as public services, contracts, accommodations, or employment (Black, 1994, pp. 155-156; Zinn & Brethour, 1996, p. 1:1). For example, B.C.'s Human Rights Code specifically states that an employer or employment agency must not discriminate against a person because of "race, colour, ancestry, place of origin, political belief, religion, marital status, family status, physical or mental disability, sex, sexual orientation or age of that person or because that person has been convicted of a criminal or summary conviction offence that is unrelated to the employment or to the intended

---

[19] Although Andrews has been refined (e.g., R. v. Turpin, 1989, and most recently Public Service, 1999), it still provides the basic framework in which the concepts of equality and discrimination are understood (Hurley, 1998, pp. 1, 3-4, 7; see also Egan v. Canada, 1995).

[20] Compare to s. 2 of B.C.'s Human Rights Code.

employment of that person" (s. 13). The Charter is worded more generally, stating that "every individual is equal before and under the law and has the right to the equal protection and equal benefit of the law without discrimination and, in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental or physical disability" (s. 15).

A prohibited ground, then, is a particular characteristic, such as race, sex, etc., that is enumerated in legislation. By implication, if a characteristic is not enumerated, it may be a lawful ground upon which to discriminate. Notwithstanding, the courts have expanded the application of s. 15 of the Charter and administrative tribunals have similarly expanded federal or provincial human rights legislation by way of the principle of "non-enumerated grounds analogous to enumerated grounds" (otherwise known as "analogous grounds"). This principle was first articulated in Andrews v. Law Society of B.C. (1989), where the court addressed discrimination against a group of persons (non-citizens) that was not enumerated in s. 15 of the Charter (i.e., not specified, as is race, sex, etc.). The court's rationale for including a non-enumerated group was that the purpose of s. 15 was to protect all groups who may suffer from discrimination (Funston & Meehan, 1994, p. 187), which can only be determined in the changing context of the "social, political, and legal fabric of our society."[21] Specifically, the court stated:

> For example, Stone J. writing in 1938, was concerned with religious, national and racial minorities. In enumerating the specific grounds in s. 15, the framers of the Charter embraced these concerns in 1982 but also addressed themselves to the difficulties experienced by the disadvantaged on the grounds of ethnic origin, colour, sex, age and physical and mental disability. It can be anticipated that the discrete and insular minorities of tomorrow will include groups not recognized as

[21] See Sheppard (1993), citing Edmonton Journal v. Alberta (Attorney General) (1989, p. 3), and Hurley, (1998), citing R. v. Turpin, (1989, p. 4) for a further analysis of the importance of "context" in the interpretation of what constitutes discrimination.

> such today. It is consistent with the constitutional status of s. 15 that it be interpreted with sufficient flexibility to ensure the "unremitting protection" of equality rights in the years to come.
>
> While I have emphasized that non-citizens are, in my view, an analogous group to those specifically enumerated in s. 15 and, as such, are entitled to the protection of the section, I agree with my colleague that it is not necessary in this case to determine what limit, if any, there is on the grounds covered by s. 15 and I do not do so. (pp. 152-153)

The analogous grounds approach defined in Andrews has been confirmed by the Supreme Court of Canada in a number of recent cases (e.g., R. v. Turpin, 1989; Egan v. Canada, 1995; Eaton v. Brant County Board of Education, 1997; and most recently, Vriend v. Alberta, 1998). Notably, the courts have stated that "analogous grounds cannot be restricted to historically disadvantaged groups if the Charter is to retain future relevance" (Hurley, 1998, p. 8, citing Vriend; see also Eldridge v. British Columbia (Attorney General), 1997).

Black (1994) is of the opinion that the Supreme Court of Canada has included analogous grounds because of the Charter's general rule in s. 15 against discrimination (i.e., "every individual is equal before and under the law and has the right to the equal protection and equal benefit of the law without discrimination and, in particular...) (p. 163). However, because provincial legislation is very specific regarding what constitutes prohibited grounds, Black suggests its protection may be deficient, although the Supreme Court of Canada has been quick to "read into" deficient legislation that is held either "underinclusive"[22] or inconsistent with the Charter. For example, in Vriend, the court held that Alberta's human rights legislation (Individual's Rights Protection Act, 1996)

[22] See Sheppard (1993) for one viewpoint on "underinclusiveness" (pp. 59-60). This is a confusing and contentious area in law, where it appears that the courts are taking upon themselves the mantle of Parliament and writing new law rather than interpreting existing law.

was underinclusive and so infringed s. 15 of the Charter because it did not include sexual orientation as a prohibited ground. As a remedy, the court held that the words "sexual orientation" should be read into the legislation.

Prior to the Public Service (1999) decision, the Supreme Court of Canada had defined three different types of discrimination: (1) direct, (2) adverse effect, and (3) systemic. Historically, these distinctions were important because the type or classification of discrimination triggered different responsibilities and remedies in the employment context (e.g., Alberta (Human Rights Commission) v. Central Alberta Dairy Pool, 1990, p. 506). In terms of legal remedies these distinctions now have little legal importance; however, a discussion of them is still important to understanding the concept of discrimination and the Supreme Court's new unified approach.

Direct Discrimination

According to the decision of O'Malley v. Simpson-Sears Ltd. (1985), direct discrimination occurs in an employment context when an employer or employment agency differentiates or makes distinctions, whether based on a policy standard or not, on the basis of a legislatively prohibited ground:

> Direct discrimination occurs ... where an employer adopts a practice or rule which on its face discriminates on a prohibited ground. For example, 'No Catholics or no women or no blacks employed here.' (p. 551)

Another example of direct discrimination would be where an employer requires employees to retire at age 60 years, which on its face is a legislatively prohibited ground as it is discrimination based on age (e.g., Large v Stratford (City), 1995, as it was applied to police officers; and Saskatchewan (Human Rights Commission) v. Saskatoon, 1989, as it was applied to fire fighters).

Adverse Effect Discrimination

Adverse effect discrimination, also known by other terminology such as adverse impact discrimination and indirect discrimination, occurs when an employer has a rule, practice, or policy which is applied equally to all persons but has the effect, whether intentional or not, of discriminating against an individual or group based on a prohibited ground. In such a case, an individual or group is the victim of discrimination due to a special characteristic (e.g., wearing a turban by baptized Sikhs) linked to a prohibited ground (e.g., religion) that is protected by law.

In O'Malley (1985), Mr. Justice McIntyre explained the legal history of adverse effect discrimination:

> The idea of treating as discriminatory regulations and rules not discriminatory on their face but which have a discriminatory effect, sometimes termed adverse effect discrimination, is of American origin and is usually said to have been introduced in the Duke Power[23] case ... in the Supreme Court of the United States. In that case, the employer required as a condition of employment or advancement in employment the production of a high school diploma or the passing of an intelligence test. The requirement applied equally to all employees but had the effect of excluding from employment a much higher proportion of black applicants than white. It was found that the requirements were not related to performance on the job, and the Supreme Court of the United States held them to be discriminatory because of their disproportionate effect on the black population. There was no provision in the relevant statute ... which directed such an interpretation.... (p. 550)

Coming to the same conclusion as the Supreme Court of the United States, Mr. Justice McIntyre in O'Malley found that adverse impact, when linked to a legally protected ground, was in fact discriminatory:

> For essentially the same reasons that led to the conclusion that an intent to discriminate was not required as an element of discrimination contravening the [Human Rights] Code I am of the opinion that this Court may consider adverse impact discrimination as described in these reasons a contradiction of the terms of

[23] Referenced as Griggs v. Duke Power Co. (1971).

> the Code.[24] An employment rule honestly made for sound economic or business reasons, equally applicable to all whom it is intended to apply, may yet be discriminatory if it affects a person or groups of persons differently from others to whom it may apply. (p. 551)

An equally famous case of adverse effect discrimination decided by the Supreme Court of Canada was that of Bhinder v. C.N. (1985), where the employer, for safety reasons, introduced a policy that required all maintenance electricians to wear a hard hat. This, on its face, appeared neutral because it applied equally to all employees. However, Bhinder, the complainant in the case, was Sikh and so was forbidden to wear anything but a turban on his head. When Bhinder refused to comply with the new policy, he was eventually terminated from his employment. The adverse, unintentional impact of the hard hat rule, then, was that the rule imposed a penalty on adherents of the Sikh religion that was not imposed upon other employees. Because discrimination based on religion is prohibited by law, and because a link was made between the prohibited ground and the complainant, the hard hat rule was therefore discriminatory. Although the Supreme Court did not rule in Bhinder's favor, the narrow issue here was whether the employer had a duty to accommodate, which will be discussed in detail later in this chapter.[25]

Systemic Discrimination

Systemic discrimination, which arises at the aggregate level, is conceptually similar to adverse effect discrimination; for example, on its face, it may not appear discriminatory and may also be unintentional (Public Service, 1999, para. 39; see also Black, 1994, p. 10; Sheppard, 1993, p. 7; and Zinn & Brethour, 1996, p. 1:6). According

---

[24] Similar to an absolute liability offence, where intent need not be proven (R. v. Sault Ste. Marie, 1978; see Verdun Jones, 1997, for a good discussion on absolute liability).

[25] In Alberta (Human Rights Commission) v. Central Alberta Dairy Pool, 1990, the Supreme Court of Canada found that it had "erred" in the way Bhinder was decided, and modified the law on the duty to accommodate, notwithstanding a bona fide occupational requirement, in cases where adverse effect discrimination is established.

to Zinn and Brethour, "Systemic discrimination arises out of long-standing stereotypes and value assumptions which create the discriminatory effect" (p. 1:6). As a result, systemic discrimination is distinguished from adverse effect discrimination only in the subtlety of effect and the fact that it is revealed at a group level.

According to a ruling by the Canadian Human Rights Commission (P.S.A.C. v. Canada (Treasury Board), 1991), the concept of systemic discrimination "recognizes that long-standing social and cultural mores carry within them value assumptions that are substantially or entirely hidden and unconscious" (p. D/349, cited by Zinn and Brethour, 1996, p. 1:6). For example, it may be apparent that a rule that requires all employees to wear a hard hat or risk losing their employment will have an adverse effect on employees who are Sikhs; however, it may not be apparent that certain types of work historically performed by women have been undervalued compared to work performed by men.

According to Zinn and Brethour (1996), the leading case defining systemic discrimination is Canadian National Railway Co. v Canada (Canadian Human Rights Commission) (1987), decided by the Supreme Court of Canada (p. 1:7). Here, citing Abella in her Report of the Commission of Equality in Employment (1984, pp. 9-10), Mr. Justice Dickson described systemic discrimination[26] in an employment context as follows:

> Systemic discrimination in an employment context is discrimination that results from the simple operation of established procedures of recruitment, hiring and promotion, none of which is necessarily designed to promote discrimination. The discrimination is then reinforced by the very exclusion of the disadvantaged group because the exclusion fosters the belief, both within and outside the group, that the exclusion is the result of "natural" forces, for example, that women "just can't do the job." (p. 210)
>
> ....

[26] Compare with O'Malley (1985), in which Mr. Justice McIntyre explained the legal history of adverse effect discrimination.

> I have already stressed that systemic discrimination is often unintentional. It results from the application of established practices and policies that, in effect, have a negative impact upon the hiring and advancement prospects of a particular group. It is compounded by the attitudes of managers and co-workers who accept stereotyped visions of the skills and "proper role" of the affected group, visions which lead to the firmly held conviction that members of that group are incapable of doing a particular job, even when that conclusion is objectively false.[27] (p. 213)

Historically, as indicated above, it has been assumed that systemic discrimination was often unintentional, but in Public Service (1999) Madam Justice McLachlin[28] commented that systemic discrimination "is now much more prevalent than the cruder brand of openly direct discrimination" (para. 29). Because unscrupulous employers couch intentional systemic discrimination in neutral language, it has avoided the legal scrutiny received by direct discrimination. The problem, though, is in proving systemic discrimination, which will be discussed later in this chapter.

Exceptions to the Discrimination Rule

No right is absolute, and the law permits discrimination (i.e., legally justified discrimination) under certain circumstances. For example, s. 1 of the Charter states that rights are subject to reasonable limits, provided such limits can be "demonstrably justified in a free and democratic society" (rf. R. v. Oakes, 1986). Section 1 of B.C.'s Human Rights Code permits discrimination[29] against persons under the age of 19 years, and s. 41 exempts discrimination by, inter alia, charitable, educational, or religious non profit organizations if their "primary purpose" is the "promotion of the interests and

---

[27] In this case, the comments by Dickson on systemic discrimination were obiter dicta. The issue under appeal was whether the Canadian Human Rights Tribunal, pursuant to the Canadian Human Rights Act, had the authority to order a remedial affirmative action program (which was found to be lawful).

[28] Now Chief Justice.

[29] Authors such as Black (1994, p. 179) and Sheppard (1993, p. 2) argue that human rights legislation does not exempt certain discriminatory practices, but rather sets out legislative policy for achieving the central goal of equality.

welfare of an identifiable group or class of persons characterized by a physical or mental disability or by a common race, religion, age, sex, marital status, political belief, colour, ancestry or place of origin".[30]

However, the Supreme Court of Canada has "repeatedly stressed the importance of interpreting human rights legislation in a broad, liberal and purposive manner in order to advance the broad policy considerations underlying such legislation" (Lovett, 1997, p. 2:1:01; see also Sheppard, 1993, p. 3). In order to advance these policy considerations, the Supreme Court of Canada has interpreted prohibiting provisions found in human rights legislation broadly and, conversely, has interpreted exemptions narrowly (Brossard (Town) v. Quebec (Commission des droits de la personne), 1988, p. 307; Gould v. Yukon Order of Pioneers, 1996, pp. 585, 601; Public Service, 1999, paras. 38, 43-44; University of B.C. v. Berg, 1993, p. 370; and Zurich Insurance Ltd., 1992, p. 339).

Within this framework, two important discrimination exceptions exist in the employment context: an equity program and a bona fide occupational requirement (BFOR).[31] After discussing the concept of equity, this section will explain the principles of a BFOR followed by an historical analysis of this defence with respect to the "duty to accommodate."

Equity Programs and the Concept of Equality

Similar to discrimination, the concept of equality is "complex and elusive" (Funston & Meehan, 1994, p. 186, citing Andrews v. Law Society of B.C., 1989, at p. 164), and nowhere is this more evident than in the attempt to reconcile the application of

[30] The application of human rights legislation to private clubs, etc., is contentious, and the Supreme Court of Canada has blurred the distinction between what are "private" or "public" clubs or social organizations (Black, 1994, p. 157, citing University of British Columbia v. Berg, 1993).

[31] These defences are identical to those found in U.S. jurisprudence (rf. Paetzold & Willborn, 1999, §1.05).

equity[32] programs with the "essence of equality."[33] Paradoxically, treating people equally does not always result in equality; i.e., what is equitable, just, and fair (Black, 1994, p. 80; Sheppard, 1993, pp. 4-5; Weiner, 1993, p. 17). Four years before Andrews, the concept of equity was discussed by the Saskatchewan Court of Appeal in Saskatchewan (Human Rights Commission) v. Canadian Odeon Theatres Ltd (1985): "The treatment of a person differently from others may or may not amount to discrimination just as treating people equally is not determinative of the issue" (p. 115). The most obvious example is that of adverse effect discrimination, where on its face a practice may appear to treat everyone equally but the effect is discriminatory against a particular group (e.g., the hard hat rule in Bhinder). As noted by Black (1994), commenting on the decision by the Supreme Court of Canada in Andrews, the principle of equality is measured in terms of effects or results (a substantive approach) rather than identical treatment (a formal approach[34]) (pp. 64-65, 80, 154, 178).[35]

Therefore, when addressing adverse effect and systemic discrimination, judges and lawmakers must necessarily apply equity principles (Sheppard, 1993, pp. iii, 10). Operationally, in the employment context, this often requires a set of activities or programs the goals of which are (1) to ensure that those groups who have traditionally

---

[32] Historically, in England and Canada various courts operated under extremely rigid rules, such as the courts of common law, which had jurisdiction to provide remedies only in very defined circumstances. If a set of circumstances were beyond the jurisdiction of the common courts, claimants could apply to a "court of equity," otherwise known as a "court of chancery," for an "extraordinary" remedy, which became known as "equitable" remedies (Gall, 1995, p. 56).

[33] According to the Supreme Court of Canada, the "essence of equality" is treating each person "according to one's own merit, capabilities and circumstances" (Public Service, 1999, para. 81).

[34] When the British Columbia Court of Appeal decided the Andrews case, it used a formal approach, where persons similarly situated were entitled to similar treatment; and where persons were differently situated, different treatment was justified. However, the Supreme Court of Canada rejected the formal approach, endorsing the substantive approach (Hurley, 1998, p. 2).

[35] For more discussion, refer to Hurley (1993, p. 2), Sheppard (1993, pp. 4-5), and Zinn & Brethour (1996, p. 1:5), and compare their comments with the court decisions of O'Malley (1985) and R. v. Turpin (1989).

been the object of discriminatory practices have an equal opportunity to compete in the job market, and (2) to ameliorate the historical negative consequences of discriminatory practices against such groups. In order to achieve such goals, substantive equality must be assessed in the "larger social, political and legal context" (R. v. Turpin, 1989; see also Black, 1994, pp. 80, 155; and Hurley, 1998, pp. 3-4).

As a result, however, equity programs can result in what is popularly known as reverse discrimination, where an individual's merits may have to yield to utilitarian goals. For this reason, affirmative action programs are often contentious (e.g., see Sheppard (1993, pp. iii, 8) and Black (1994, p. 155), citing Action Travail des Femmes v. C.N. Railway (1987) in which the Supreme Court of Canada directed the employer to implement an affirmative action program). Nevertheless, human rights legislation specifically exempts such programs from allegations of discrimination. For example, s. 42 of B.C.'s Human Rights Code states:

> (1) It is not discrimination or a contravention of this Code to plan, advertise, adopt or implement an employment equity program that
> (a) has as its objective the amelioration of conditions of disadvantaged individuals or groups who are disadvantaged because of race, colour, ancestry, place of origin, physical or mental disability, or sex, and
> (b) achieves or is reasonably likely to achieve that objective.

Similarly, s. 15 of the Charter states:

> (1) Every individual is equal before and under the law and has the right to the equal protection and equal benefit of the law without discrimination and, in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental or physical disability.
> (2) Subsection (1) does not preclude any law, program or activity that has as its object the amelioration of conditions of disadvantaged individuals or groups including those that are disadvantaged because of race, national or ethnic origin, colour, religion, sex, age or mental or physical disability.

Here, Weiner (1993) argues that employment equity "is not about giving ... groups an advantage, but to provide them with their fair share of employment opportunities by overcoming the effects of past and present discrimination" (p. 1). Employment equity, then, is not about hiring unqualified applicants:

> There may be many reasons why designated group members have not traditionally been hired into work organizations, or have not been hired into decision-making jobs which have nothing to do with their qualifications. So, rather than hiring unqualified people, ... [employment equity] is about ensuring that job qualifications are truly job-related, the largest pool of applicants possible is generated, and everything is done to ensure that talent and qualifications are recognized, even when they are "packaged" differently [emphasis in original]. (Weiner, p. 9)

Employment equity is about a workforce that proportionally represents women, native Indians, disabled persons, visible minorities, and other protected groups, the only qualifiers being bona fide ability and qualifications, discussed next.

Bona Fide Occupational Requirement (BFOR)[36]

Human rights legislation generally provides an opportunity for an employer or employment agency to demonstrate that a discriminatory hiring or selection practice is justified because of a BFOR (e.g., s. 13(4) of B.C.'s Human Rights Code). A BFOR is a rational link between an occupation (e.g., police officer) and an organizational policy or workplace standard (e.g., vision requirement) that discriminates on a legislatively prohibited ground (e.g., physical disability).[37] Here, the employer has demonstrated on a

[36] Also known by other terms such as "bona fide occupational qualification" (BFOQ). As noted in the introduction to this chapter, discrimination in the employment context is similar to that in other contexts. For example, a BFOR is similar to a "bona fide and reasonable justification" (BFRJ), which refers to justified discriminatory practices with respect to accommodations and services (rf. s. 8 of B.C.'s Human Rights Code, and the case British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights, 1997).

[37] The difference between the legal concept of a BFOR and the psychometric concept of test validity will be discussed later in this chapter.

"balance of probabilities"[38] that the discriminatory policy or standard is necessary in order to select employees who can safely and competently perform the necessary tasks involved in the occupation. In 1989, Mr. Justice Sopinka, in Saskatchewan (Human Rights Commission) v. Saskatoon, defined the concept of BFOR as follows:

> The general philosophy of human rights legislation is that persons are not to be judged or dealt with on the basis of external characteristics such as race, age, sex, etc. but on individual merit. That is the general rule and violation of it constitutes discrimination... The defence of a bona fide occupational qualification or requirement is an exception to the general rule. In the limited circumstances in which this defence applies, it is not individual characteristics that are determinative but general characteristics reasonably applied. (pp. 1309-1310)

At that time, if the employer made out a BFOR case, direct discrimination against an individual or group was permissible and there was no concomitant duty to accommodate an individual (discussed more fully below, under the heading "Duty to Accommodate"). The rationale for this exemption was that, logically, accommodation is inconsistent with the concept of the defence, as noted in Alberta (Human Rights Commission) v. Central Alberta Dairy Pool (1990):

> Either it is valid to make a rule that generalizes about members of a group or it is not. By their very nature rules that discriminate directly impose a burden on all persons who fall within them. If they can be justified at all they must be justified in their general application. (p. 514)

In evaluating a BFOR defence, the Supreme Court of Canada had constructed a two-part analysis: a subjective test and an objective test. The subjective test referred to what the employer honestly believed, its purpose to "ensure that a discriminatory rule was adopted for a valid reason" (Lovett, 1997, p. 2:1:15; rf. Large v Stratford (City), 1995, pp. 745-746). In other words, a valid reason did not include an intention to subvert

[38] Rf. Ontario Human Rights Commission v. Etobicoke, 1982, p. 503.

the legislation; rather, it was one held "honestly, in good faith; and in the sincerely held belief that such limitation ... [was] imposed in the interests of the adequate performance of the work involved with all reasonable dispatch, safety, and economy; and nor for ulterior or extraneous reasons [emphasis in original]" (Large, pp. 744-745).

The objective test referred to the evidence[39] upon which the employer's belief was based, its purpose to demonstrate that the discriminatory standard, policy, or rule was reasonably necessary (Saskatchewan (Human Rights Commission) v. Saskatoon [Firefighters], 1989, p. 1310). According to Ontario Human Rights Commission v. Etobicoke (1982):

> [The standard, policy, or rule must] be related in an objective sense to the performance of the employment concerned, in that it is reasonably necessary to assure the efficient and economical performance of the job without endangering the employee, his fellow employees and the general public. (p. 208)

In issues of safety, the test was that of "sufficient risk," which was less than a "substantial risk" (as set out in Etobicoke, 1982) but greater than a "minimal risk" (British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights, 1997, citing Canada (Human Rights Commission) and Husband v. Canada (Armed Forces), 1994).

As noted in Saskatchewan (1989) above, the BFOR defence had a general application based on the average characteristics of the group to which it applied. Notwithstanding, there was a duty on the employer to show that no reasonable alternative existed (Alberta (Human Rights Commission) v. Central Alberta Dairy Pool, 1990, p.

---

[39] To meet the objective test, the evidence does not necessarily have to be empirical, but must be more than impressionistic or speculative (rf. British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights (1997), where the court compared Ontario Human Rights Commission v. Etobicoke (1982), which failed the objective test, to Saskatchewan (Human Rights Commission) v. Saskatoon [Firefighters] (1989), which passed the objective test (both cases addressing the issue of mandatory retirement for firefighters at age 60 years)).

518). In other words, the employer was required to show that to test each applicant or employee on an individual basis was not reasonable (Large v Stratford (City), 1995, pp. 349, 751; Ontario Human Rights Commission v. Etobicoke (1982), pp. 209-210; Saskatchewan (Human Rights Commission) v. Saskatoon [Firefighters], 1989, p. 519; and Zurich Insurance Ltd. v. Ontario (Human Rights Commission), 1992, p. 349).

Duty to Accommodate

Historically (prior to Public Service, 1999), in order to assess whether there was a "duty to accommodate" within a BFOR argument, it was first necessary to distinguish between direct and adverse effect discrimination. The concept of duty to accommodate is not found in human rights legislation, but was developed by the courts and adopted by the Supreme Court of Canada in Ontario Human Rights Commission v. Etobicoke (1982), and applied in the context of adverse effect discrimination only (Alberta (Human Rights Commission) v. Central Alberta Dairy Pool (1990), p. 513). In Bhinder v. C.N. (1985) and O'Malley v. Simpson-Sears Ltd. (1985), both of which were decided simultaneously, the Supreme Court of Canada addressed the concept of duty to accommodate as it applied to direct and adverse effect discrimination.

In O'Malley, the complainant was sometimes required to work on Friday evenings and Saturdays. After working for her employer for approximately three years, the complainant became a "Seventh-Day Adventist," a religious denomination that strictly observes the Sabbath from Friday evening to Saturday evening. As a result, the complainant requested that she be exempted from working Friday evenings and Saturdays. O'Malley was distinguished from the facts of Bhinder only in that the relevant human rights legislation did not contain a provision for BFOR. As a result,

because the employer in O'Malley could not claim this defence, the court ruled that the employer had a duty to accommodate the complainant (up to a point of undue hardship) because the case was one of adverse effect discrimination. However, in Bhinder, because the relevant human rights legislation contained a provision for BFOR, notwithstanding the case was one of adverse effect discrimination, the court ruled that the employer had no duty to accommodate the complainant because the criteria for BFOR were satisfied (p. 590).

The consequence of the O'Malley and Bhinder decisions was that if the relevant legislation provided for a BFOR defence, accommodation was not required in cases of either direct or adverse effect discrimination. The logic was that a successful BFOR defence was not compatible with an individualized approach; that is, an "occupational" requirement vis-a-vis an "individual" assessment (Lovett, 1997, p. 2:1:10). These decisions were sharply criticized as they allowed adverse effect discrimination (Lovett, p. 2:1:10). Philosophically, in cases of direct discrimination the BFOR principle was considered morally acceptable because it was job-related and its application was truly "general." On the other hand, in cases of adverse effect discrimination, the application was not genuinely general—a burden (e.g., loss of employment) not imposed upon the main group (e.g., all electricians) was imposed upon a sub-group (e.g., Sikh electricians) on a legislatively prohibited ground (e.g., religion).

Confronted with such well argued criticism, in Alberta (Human Rights Commission) v. Central Alberta Dairy Pool (1990) the Supreme Court of Canada seized on an opportunity to re-visit the principle established in O'Malley and Bhinder. The facts in Alberta were similar to O'Malley and Bhinder, allowing the court to modify the law

regarding the application of a BFOR defence. Specifically, the issue was that of adverse effect discrimination where the complainant, because of religious convictions, could not work certain holy days and as a result was dismissed by his employer. Moreover, as in Bhinder, the relevant human rights legislation provided for a BFOR defence. Madame Justice Wilson, writing for the majority, rewrote the law as follows:

> My ... reason for questioning the correctness of Bhinder concerns the assumption that underlies both the majority and minority judgments, namely that a BFOR defence applies to cases of adverse effect discrimination. Upon reflection, I think that we may have erred in failing to critically examine the assumption. As McIntyre J. notes in O'Malley, the BFOQ test in Etobicoke was formulated in the context of a case of direct discrimination on the basis of age. The essence of direct discrimination in employment is the making of a rule that generalizes about a person's ability to perform a job based on membership in a group through sharing a personal attribute such as age, sex, religion, etc. The ideal of human rights legislation is that each person be accorded equal treatment as an individual taking into account those attributes. Thus, justification of a rule manifesting a group stereotype depends on the validity of the generalization and/or the impossibility[40] of making individualized assessments [emphasis added]. (p. 513)

As a result of this decision, it was necessary to take a hybrid approach to discrimination. In cases of direct discrimination the employer was obligated to establish that the workplace standard was a valid BFOR by showing (1) that the standard was implemented honestly and in good faith without unlawful ulterior motives, and (2) that the standard was objectively necessary in order to perform the work safely and efficiently. If the employer failed, the standard was struck down and accommodation was therefore not relevant. On the other hand, if the employer succeeded, the standard was upheld and the employer was under no general duty to accommodate. In cases of

---

[40] In Large v Stratford (City) (1995, at p. 751), the Supreme Court of Canada interpreted "impossibility" to mean "impracticality" (rf. British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights, 1997). However, in Public Service (1999), the term "impossible" was again introduced, which suggests a return to a very strict application of accommodation (para. 54).

adverse effect discrimination, the employer was obligated to show (1) that a rational connection existed between the workplace standard and the job, and (2) that accommodating[41] an individual[42] adversely affected was not possible without incurring undue hardship.[43] If the employer failed, the individual succeeded in a claim for accommodation, but notably the workplace standard itself remained unchallenged.

In cases of systemic discrimination the ameliorative advantages of this hybrid approach were unclear. In direct systemic discrimination (e.g., sex), if the workplace standard were not a valid BFOR, the remedy allowed for an affirmative action (equity) program in addition to that of an individual remedy. For example, in the case of Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987), the Supreme Court of Canada ruled that the employer must implement an affirmative action hiring program for women to remedy the historical practice of excluding women from working in jobs traditionally held by men. However, if the workplace standard were a valid BFOR, although it discriminated against women, accommodation was theoretically not available. In adverse effect discrimination, although accommodation was theoretically available, the question was how systemic discrimination could effectively be addressed if the offending workplace standard were left unchallenged. This problem is discussed next.

---

[41] For a short legal history on the duty to accommodate, see O'Malley v. Simpson-Sears Ltd. (1985).
[42] The duty to accommodate extends also to the employee and any representative union (rf. Central Okanagan School District No. 23 v. Renaud, 1992, pp. 992-995; 986-87; Chambly, Com'n Scolaire v. Berguein, 1994, p. 627; and O'Malley v. Simpson-Sears Ltd., 1985, p. 555).
[43] For a discussion on what constitutes "undue hardship," see Alberta (Human Rights Commission) v. Central Alberta Dairy Pool (1990, pp. 514-516, 520-521), Central Okanagan School District No. 23 v. Renaud (1992, pp. 984-985), and Chambly, Com'n Scolaire v. Berguein (1994, p. 626).

Discrimination Redefined

From the preceding discussion it is apparent that the Supreme Court of Canada has labored greatly and amended its own decisions on a number of occasions in an attempt to refine the fundamental principles underlying the concept of discrimination. To follow complex and shifting legal principles is difficult enough, but for employers or employment agencies to apply those principles operationally is a formidable challenge. Take for example the recent case in which the British Columbia Court of Appeal (British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union, 1997) overturned an arbitrator's decision that required an employer to accommodate an employee on the basis that adverse effect discrimination had occurred (rf. Lovett, pp. 2:1:28-29).

Problems of Practical Application

This case involved a woman (Tawney Meiorin) who was employed for three years by the Ministry of Forests as a member of a first response fire fighting crew. The work is physically demanding and potentially dangerous, and so the employer, in response to a 1991 Coroner's Inquest Report, created a fitness standard. In 1994, Meiorin was required to take the fitness test, but after four attempts she failed to meet the aerobic standard (a running test). Despite the fact that she had passed the other components of the fitness test and had previously received satisfactory work reports, Meiorin's employment was terminated.

In response, Meiorin's union grieved her termination to an arbitrator, who heard evidence that the fitness test's aerobic standard was so demanding that only 35% of women passed the standard compared to 65% of men. On its face, the aerobic fitness

standard appeared neutral—all employees had to meet the standard in order to meet their employment obligations. However, because on average men have a greater aerobic capacity than do women, it was argued that the standard had an adverse impact against women, as indicated in their disproportionate failure rate. The arbitrator agreed, ruling that Meiorin had established a prima facie case of adverse effect discrimination and that the employer had failed to accommodate her as required by law. As a result, because no evidence was presented that indicated Meiorin was a safety risk, the arbitrator ordered her reinstated, a decision the employer appealed to the British Columbia Court of Appeal.

The Court of Appeal overruled the arbitrator, finding that the fitness standard was a BFOR and therefore did not constitute discrimination on the prohibited ground of sex. Interestingly, the court did not address the issue of direct vis-a-vis adverse effect discrimination; rather, it reasoned that because the distinction was not based on personal characteristics attributed to an individual solely on the basis of association with a group, but rather based on individual merits and capabilities relevant to the occupation, discrimination did not occur (rf. Andrews v. Law Society of B.C., 1989, pp. 174-175). Additionally, the court reasoned that because the complainant was individually tested, discrimination could not have occurred (rf. Large v. v Stratford (City), 1995). Finally, the court reasoned that should the fitness standard be lowered to allow the same proportion of women to pass as men, then "a new concept which could be labelled 'reverse/adverse effect discrimination'" would be introduced. Here, men who could meet the lower standard for women but not the higher standard for men would be automatically excluded from employment simply because of being male.

On appeal by the union, this decision was subsequently overturned by the Supreme Court of Canada (Public Service, 1999) on the grounds that the researchers' methodology was flawed and that the aerobic standard was not a valid BFOR (the issue on which Court of Appeal was found to be mistaken). Although Meiorin was ordered reinstated as a fire fighter, the rationale of the Supreme Court differs substantively from that of the arbitrator and marks an important new direction in the law on discrimination. In addition to ruling on the narrow issue of Meiorin's termination, the Supreme Court addressed the broader issue of systemic sex discrimination—whether the workplace standard unfairly excluded women from employment as fire fighters.

As discussed in this chapter, the conventional approach to human rights law required a distinction between direct and adverse effect discrimination, which the Supreme Court now ruled was fraught with "profound" difficulties (para. 24). For example, the distinction between direct and adverse effect discrimination was found to be artificial and vague. Because the remedy for discrimination was dependent on whether it was direct or adverse effect, the Supreme Court found that adjudicators may have artificially characterized claims in order to apply the remedy they believed was most equitable in the circumstances. Moreover, the Supreme Court found it "disconcerting" that the remedies differed, especially since one could potentially offer more protection than the other and so itself become discriminatory (paras. 31, 33).

As a result the Supreme Court found that the conventional approach was theoretically incoherent, giving rise to problems of practical application and possibly interfering with the broad purpose of human rights legislation (i.e., to promote equality, prevent discrimination, and provide an equitable means of redress) (para. 38). For

example, systemic discrimination may be unintentionally reinforced because if characterized as adverse effect, the offending workplace standard would be considered neutral and its legitimacy therefore never challenged (para. 39). Here, the focus shifts from the offending standard to that of accommodation, placing the underlying causes of systemic discrimination beyond judicial scrutiny.

Finally, the Supreme Court found that the conventional approach to human rights legislation was inconsistent with its approach to section 15 of the Charter. In Charter claims, once unjustified discriminatory effect is shown, a remedy is not dependent upon the nature of the discrimination or whether it was unintentional. As stated by McLachlin, "I see little reason for adopting a different approach when the claim is brought under human rights legislation which, while it may have a different legal orientation, is aimed at the same general wrong as s. 15(1) of the Charter" (para. 48).

A New Unified Approach

The significance of the Public Service (1999) case was that the Supreme Court went beyond striking down an invalid workplace standard to reviewing the law on discrimination, finding it deficient, and substituting a new unified approach.[44] Essentially, this unified approach eliminated the need to characterize discrimination as either direct or adverse effect (although the Court recognized some analytical value in doing so), defined all discrimination in terms of "discriminatory effect," and provided a three-step test for assessing the validity of a BFOR defence. To justify a discriminatory workplace standard, an employer must now establish the following steps on a balance of probabilities:

[44] Such an approach had been previously advanced in the literature, administrative and judicial rulings, and some human rights legislation, which was noted by the Supreme Court (paras. 50-53).

(1) the standard was selected for a legitimate purpose rationally connected to job performance, which includes, but is not limited to, safety and efficiency,[45] which are objective requirements of any job;

(2) the standard was adopted honestly and in good faith, without an unlawful ulterior motive; and

(3) the standard is reasonably necessary to accomplish the employer's legitimate work-related purpose, which can only be shown if it is "impossible to accommodate individual employees sharing the characteristic of the claimant without imposing undue hardship[46] upon the employer" (para. 54).

If the workplace standard cannot meet this test, it will be struck down, and the remedy may include any of those found in the applicable human rights legislation. Conversely, if this test is met the discriminating workplace standard will be found to be justified. In applying this test to the facts of Public Service (1999), the Supreme Court held that Meiorin had demonstrated a prima facie case of adverse effect discrimination (i.e., the aerobic standard discriminated against her on the prohibited ground of sex, because women have on average lower aerobic capacity than men). Although the employer had demonstrated a rational connection between the workplace standard and the job of firefighter, it could not show that the standard was reasonably necessary for the safe and efficient performance of the job (step 1). Moreover, although the standard had been adopted in good faith (step 2), the employer could not show that the standard was reasonably necessary for the accomplishment of the work-related purpose—i.e., that it

[45] Other reasons exist, such as preserving the integrity of religious schools, anti-nepotism policies, etc.
[46] The Supreme Court referred to its own jurisprudence for guidance on what constitutes "undue hardship," which has been previously discussed in this chapter.

was impossible to accommodate individual employees sharing the characteristics of Meiorin without undue hardship (step 3).[47]

Assessing Systemic Sex Discrimination

As discussed, all discrimination is not necessarily wrong; moreover, even when ostensibly wrong, human rights legislation permits an employer or employment agency to demonstrate that discriminatory workplace standards are justified because of a BFOR. Although job-relatedness underlies the logic of the BFOR defence, is should not be confused with the psychometric concept of test validity as generally applied by industrial psychologists in an employment context. Here, test validity focuses on the operationalization of specified constructs or behaviors empirically related to the target job for the purpose of maximizing discrimination on the dimension(s) of interest.

The objective of such tests is to predict individual performance in the target job, which on the one hand may be valid within a psychometric framework but on the other hand may be invalid within a legal human rights framework. For example, on conventional cognitive ability tests black people generally perform lower than white people (Hoffman & Thornton, 1997, p. 461, citing Gottfredson, 1988; Howard & Bray, 1988, pp. 339-342; Singer, 1993, p. 23),[48] and on physiological aerobic capacity tests

---

[47] Ironically, the Court stated that this test was a "simpler, more common sense approach" (para. 53), but confused the steps of the test in its application to Meiorin. The Court specifically found that the employer had met the first two steps, but when addressing step three the Court found that the employer had not demonstrated that the workplace standard was necessary for the safe and efficient performance of fire fighting (paras. 73-75). The Court, here, appears to contradict itself as this criterion was previously described by the Court as belonging to step one (paras. 54, 57-59), while step three was to assess the reasonableness of the standard against the employer's duty to accommodate (paras. 54; 62-68).

[48] Critics argue that conventional intelligence tests do not measure mental ability as much as they do middle-class values, which explains the higher scores by white people compared to black people (e.g., see Schmidt, Berner & Hunter, 1973).

women generally perform lower than men (Public Service, 1999). Although an employer may argue, within a utility framework, that such tests allow for the most economical selection of the highest performing candidates for the target job, such an approach has been rejected in law as wrong. For example, in Public Service, Madame Justice McLachlin accepted Day and Brodsky's (1996) criticism of the conventional approach to accommodation in the workplace:

> The difficulty with this paradigm is that it does not challenge the imbalances of power, or the discourses of dominance, such as racism, able-bodyism and sexism, which result in a society being designed well for some and not for others. It allows those who consider themselves "normal" to continue to construct institutions and relations in their image, as long as others, when they challenge this construction are "accommodated". (para. 41)

Consistent with its substantive approach to equality (where the focus is on effects), the Supreme Court's new unified approach demonstrates a greater concern for the validity of the more general purpose of workplace standards, while the conventional approach focussed more on the traditional issue of job performance (Public Service, 1999, para. 59). Although discrimination may be necessary for the safe and efficient performance of a particular job, employers must now be more aware of the differences between individuals and the differences that characterize groups of individuals for the purpose of building conceptions of substantive equality into workplace standards (para. 68).

For example, in Public Service (1999) the Supreme Court criticized researchers for not distinguishing between males and females when constructing the workplace standard for fire fighters, notwithstanding testimony by an expert who defended the methodology as valid in the conventional sense (paras. 10, 73, 74). The Supreme Court stated, "While the researchers' goal was admirable, their aerobic standard was developed

through a process that failed to address the possibility that it may discriminate unnecessarily on one or more prohibited grounds, particularly sex. This phenomenon is not unique to the procedures taken towards identifying occupational qualifications ...[emphasis added]" (para. 75). The Court went on to warn that "employers and researchers should be highly mindful of this serious problem" (para. 75).

The Significance of "Effects"

In order to advance equality, discrimination must be both identified and prevented, which is especially difficult in cases of systemic discrimination. As noted by Zinn and Brethour (1996), "One of the most obvious problems in bringing a systemic discrimination complaint is finding sufficient evidence to support it" (p. 1:8). However, since the Supreme Court adopted the "effects theory" where intent is not an issue and the focus is on the results (Bhinder v. C.N., 1985; O'Malley v. Simpson-Sears Ltd., 1985),[49] the utility of statistical evidence has assumed a new and important role in discrimination litigation.

For example, in Lasani v. Ontario (Ministry of Community and Social Services (1993) the Canadian Human Rights Board of Inquiry said that evidence of systematic discrimination generally includes not only documentation of the attitudes of supervisors, incidents, etc., but also a statistical analysis for the purpose of making comparisons (para. 49; see also Hurley, 1987, pp. 27-28; and Sheppard, 1993, p. 11).[50] Specifically, the Board noted that this mixed methodology "was the mode of analysis approved by the

[49] Under the Supreme Court's new unified approach, all discrimination is now defined in terms of "discriminatory effect" (Public Service, 1999).

[50] See Balzer's (1976) article entitled, "A View of the Quota System in the San Francisco Police Department," for one of the first published examples in policing when a statistical analysis was used to demonstrate a prima facie case of discrimination.

Supreme Court of Canada in C.N.R. [Canadian National Railway Company] v. Canada (Canadian Human Rights Commission)" (D/421).[51] Regarding the relevance of statistical evidence, the Board in Lasani stated the following:

> Overall, the evidence is clear that the Ministry has hired minorities at a rate congruent with what would be expected given their general representation in the Hamilton area. While such statistical information can seldom be determinative in a given instance, a person who asserts ethnic discrimination will be helped in cases in which there is a significantly significant disequilibrium between the number of minority candidates hired and their representation in the general population. By the same token, a person who asserts that he has been refused promotion because of racial or ethnic factors may have difficulty proving the relationship to the prohibited ground when other members of his racial minority have apparently not faced similar difficulties in obtaining promotion.

A comparative analysis is essential to the concept of equality and logically in any determination of discrimination, systemic or otherwise, as explained in Battlefords & District Co-operative v. Gibbs (1996):

> A finding of discrimination based on the imposition of a burden or the withholding of a benefit must be rooted in a comparison of the treatment received by a person with the treatment received by other persons. As McIntyre J. stated in Andrews, supra, at p. 164:
>
> > The concept of equality has long been a feature of Western thought ... It is a comparative concept, the condition of which may only be attained or discerned by comparison with the condition of others in the social and political setting in which the question arises [emphasis added]. (p. 585)

Moreover, discrimination "on the basis of association with a group" includes a sub-set of that group, and may be determined by comparing the treatment of the sub-set with the remainder of the group (Lovett, 1997, p. 2:1:02).

---

[51] Note the similarity to the seminal case of McDonnell-Douglas Corp. v. Green, heard by the Supreme Court of the United States in 1973. This case was one of the first in the United States to provide a methodology for assessing discrimination, which established that the plaintiff could rely on anecdotal, comparative, and statistical evidence (rf. Paetzold & Willborn, 1999, § 3.01).

Given the focus on effects and comparisons, in cases of systemic discrimination, statistical evidence (which, technically, is circumstantial or indirect evidence) is often crucial because it can identify and quantify patterns of behavior involving groups that are not apparent to casual observation.[52] According to Vining, McPhillips and Boardman (1986), "In most current cases, establishing employment discrimination would be impossible without the introduction of such evidence" (p. 668; see also Vizkelety, 1987, pp. 133, 174, 176).

Nevertheless, the use of statistics in discrimination cases has not received the same judicial attention as it has in the United States (U.S.), which has been noted in a number of Canadian cases.[53] In 1984, the tribunal in Blake v. The Ministry of Correctional Services noted that "statistical evidence is commonly used in American discrimination cases but has only recently been adduced at Canadian human rights hearings" (cited by Vining et al., 1986, p. 697). Also in 1984, when the Canadian Human Rights Tribunal heard the case of Action Travail des Femmes v. C.N. Railway,[54] it cited over 30 U.S. cases but only one Canadian case.[55] Notably, when this case was appealed, the Federal Court of Appeal not only approved of the practice of reviewing U.S. jurisprudence in discrimination cases but also stated that to do otherwise might be "delinquent" (rf. Vining et al., 1986, p. 662).

---

[52] See Vizkelety (1987) for a discussion on circumstantial evidence (pp. 140-144).
[53] With respect to using statistical evidence in discrimination litigation, Vizkelety (1987) described Canada as being in the "infancy stage" (p. 173).
[54] When heard by the Supreme Court, this case was indexed as Canadian National Railway Co. v. Canada (Canadian Human Rights Commission), [1987] 1 S.C.R. 1114.
[55] Similarly, when discussing the application of statistics to discrimination, this study must necessarily refer to U.S. studies and jurisprudence.

Evidence of Discrimination

Generally, in a determination of systemic discrimination in an employment context, the first question is whether there is evidence that the workplace standard, test or policy discriminates on a prohibited ground (e.g., sex, race, religion, etc.). Specifically, with respect to this study, the question is whether there is evidence to indicate that the Police Academy assessment center discriminates on the basis of sex. In order to answer this question, which legally is one of parity, it is necessary to conduct a comparative analysis, where the assessment scores of females are compared with those of males. As discussed above, the analysis can be qualitative (e.g., anecdotal evidence) or quantitative (e.g., statistical evidence). Should the results differ, it may constitute prima facie evidence of discrimination, otherwise known as "discriminatory effect."

If there is evidence of discriminatory effect (e.g., there is a systematic difference in scores), the second question is whether the difference in scores is sufficient to be legally significant, which is generally interpreted to mean practically significant.[56] In other words, the question is whether any systematic difference between the overall assessment scores of females and those of males is sufficiently large to be meaningful (or practical). In order to answer this question, which arguably is one of probity,[57] it is

---

[56] Legal significance is related to prima facie evidence and equally as difficult to define. Prima facie evidence is generally defined as minimal evidence that, if unanswered, is sufficient for a finding of fact (Yogis, 1983, p. 166, citing Girvin v. The King, 1911). Although the term legal significance does not appear in Canadian jurisprudence, it has been used in the context of discrimination by U.S. courts since 1976, and has been generally interpreted to mean practical significance (Vining et al., 1986, p., 668, citing United States v. Test, 1976). Similarly, for the purposes of this study, legal significance is generally interpreted to mean practical significance, which, as will be seen, defies any attempt at a precise definition.

[57] Form the Latin, probus, which means good, and often refers to integrity that is tested and confirmed. The application here is that the assessment model (proposed by this study) tests the integrity of personnel selection standards on the moral issue of discrimination.

necessary to measure the difference in scores against an acceptable standard, which may be qualitative (e.g., persuasive argument) or quantitative (e.g., statistical significance).[58]

In establishing a standard or benchmark for legal significance, especially as it applies to evidence in cases of discrimination, the Canadian courts have not been particularly helpful. For example, in Lasani (1993) the Board of Inquiry suggested that a "significantly significant disequilibrium" would be sufficient, while the Supreme Court in Public Service (1999) suggested a "disproportionately negative effect."[59] The U.S. jurisprudence, however, is more helpful, both in terms of case law (which as discussed above is much more developed) an in terms of regulatory law, such as the Uniform Guidelines (Equal Employment Opportunity Commission, 1978).[60] This federal

---

[58] Legal significance is conceptually similar to statistical significance, as both are inferential in nature. However, whereas statistical significance has the advantage of being determined by precise alpha levels, legal significance in the context of discrimination has the disadvantage of being determined by an ambiguous "disproportionately negative effect" (e.g., see Canadian National Railway Co. v. (Canadian Human Rights Commission), 1987, and Public Service, 1999). In summary, legal significance depends upon circumstances, fact, and law; and if statistical evidence is introduced, the question is simply whether it is material and sufficient.

[59] In Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987), the Supreme Court found a prima facie case of discrimination based upon evidence that the proportion of women hired was "substantially lower" than that found on average amongst comparable employers (0.70 compared to 13.0, respectively). For a discussion on the previous federal tribunal ruling, see also Vizkelety (1987, p. 170).

[60] The Equal Employment Opportunity Commission (EEOC), created by the federal Civil Rights Act (1964) of the United States, is the enforcement agency for that Act (Landrine & Klonoff, 1997, p. 178). In 1966, under authority of this legislation, pursuant to Title VII, which inter alia applies to discriminatory employment practices, various federal agencies, including the EEOC, adopted the Uniform Guidelines on Employee Selection Procedures, which was revised in 1970 and 1978 (Dreher & Sackett, 1981, p. 553; Hurley, 1987, pp. 29-31). As stated in the title, this regulation is a "guideline" rather than prescriptive law, although various sections may well have the force of law if underpinned by case law (Norton, 1981, p. 565; Hurley, p. 33). In 1971 the United Stated Supreme Court in Griggs v. Duke Power Co. held generally that the EEOC Guidelines should be given "great deference" (cited in Vining et al., 1986, p. 688). However, when organizations such as the International Congress on the Assessment Center Method or American Psychological Association develop professional standards, courts in the United States have given them deference, even if in conflict with the Guidelines. For example, in a case involving an assessment center (Berry v. Omaha, 1975), the judge relied heavily on the Standards and Ethical Considerations for Assessment Center Operations (1975) when finding for the employer (Hurley, pp. 33-35, 38). Notably, for enforcement purposes, the EEOC "depend heavily upon the use of comparative statistical data in the enforcement of Title VII objectives" (Hurley, p. 30).

regulation uses the "four-fifths" rule,[61] which in the U.S. is the predominant approach in cases of systemic discrimination, and has been described as an "easy-to-apply guide" that is appropriate in most cases (Paetzold & Willborn, 1999, §§ 5.05, 5.06).

To resolve the uniquely Canadian problem of determining legal significance in cases of systemic discrimination, consistent with the advice of the Federal Court of Appeal in Action Travail[62] to review the practice of the U.S., this study recommends applying U.S. methodology within a Canadian framework. The qualitative approach is used similarly in both countries and so additional discussion on this point is unnecessary, but there is divergence (not disagreement) on how the quantitative approach is used. Here, U.S. jurisprudence provides more precise standards against which discriminatory effects may be judged,[63] which may be particularly instructive in the Canadian context. Importantly, there is no suggestion that the qualitative approach, which can provide contextual data not revealed in a quantitative analysis, is unimportant just because it is not the focus of this study. Both approaches are important, but the present study focuses on the quantitative approach to demonstrate its potential for identifying discrimination, especially systemic discrimination.

---

[61] According to this rule, adverse effect is assumed when 80% of a minority group cannot meet the success rate of the dominant group on a particular standard or test (Hoffman & Thornton, 1997, p. 461), although differences of less than 20%, when significant statistically and practically, may constitute adverse effect (Hurley, 1987, p. 40). This method will be addressed in detail in Chapter 5 (Research Design).

[62] When heard by the Supreme Court, this case was indexed as Canadian National Railway Co. v. Canada (Canadian Human Rights Commission), [1987] 1 S.C.R. 1114.

[63] For example, see Hazelwood School Dist. v. United States (1977). Although this case was addressing "systemic disparate discrimination" (similar to Canada's "adverse effect" discrimination), it encompasses all "disparate impact" cases (similar to Canada's "systemic discrimination"). According to Paetzold and Willborn (1999), this was one of a number of landmark cases heard by the U.S. Supreme Court, which approved of plaintiffs using sophisticated statistical evidence to prove discrimination (§§ 3.01, 4.01, 4.08).

Finally, it should be noted that in both the U.S. and Canada, the law regarding the justification of a discriminatory practice is similar. If a case is made for a BFOR, no legal obligation exists on the employer to eliminate the discriminatory effect. On the other hand, if a workplace standard does not discriminate on a prohibited ground, no legal obligation exists on the employer to demonstrate a BFOR. In other words, in the absence of discrimination no legal obligation exists on the employer to validate a personnel selection standard (cp. Public Service, 1999, para 70, with Equal Employment Opportunity Commission, 1978; see also Dreher & Sackett, 1981, p. 558; Hoffman & Thornton, 1997, p. 456; and Hurley, 1987, p. 33). For example, an employer can quite legitimately select employees at random provided a discriminatory effect is not evident. Nevertheless, if a workplace standard does not discriminate on a prohibited ground, but evidence suggests that the standard reliably predicts relevant performance on a target job, an employer then ought to pursue it, from both a business point of view (where maximizing efficiency is a concern) and a moral point of view (where distributing benefits according to merit is a concern).

CHAPTER V: RESEARCH DESIGN

This chapter is divided into three sections. The first section provides an explanation of Police Academy assessment center operations.[1] This explanation includes a description of the participants, structure, exercises, dimensions, and scoring. The second and third sections address interrater reliability and discrimination respectively. These sections define the research questions and provide an explanation of the respective methodologies, which include descriptions of the samples, data collection procedures, and data analysis techniques.

Police Academy Recruit Assessment Center

In 1977 Development Dimensions International (DDI), under the personal direction of W.C. Byham, was hired to get the Police Academy assessment center started, and so it was not surprising to find a traditional assessment center rooted in dimension-based theory (rather the alternative exercise-based theory). Here, candidates participate in a series of job-related exercises, which are designed to elicit behavior that is classified according to pre-defined dimensions and rated according to performance standards. To understand the Police Academy assessment center, it is important to locate its operations within this theoretical and historical framework (see Chapter 2).

Participants

There are three different participants in the Police Academy recruit assessment center: the administrator, the raters, and the candidates.

[1] The historical background of the Police Academy assessment center was provided in Chapter 1, and the history and theory of the assessment center method was provided in Chapter 2.

Administrator. The assessment center administrator is an experienced police officer, usually holding the rank of sergeant, who has had extensive experience as a rater. Academy practice is to second the administrator from a municipal police department for a period of at least three or more years. The primary responsibility of the administrator is to organize operational assessment centers and to ensure that they are conducted according to the standards and guidelines established by the International Congress on the Assessment Center Method (Guidelines, 1989). This includes organizing and teaching courses on the assessment center method, providing feedback interviews to candidates, and chairing the integration session where raters share their observations and decide on an overall assessment rating for each candidate.

Raters. Similar to the practice of police assessment centers in the United States (e.g., see Fitzgerald & Quaintance, 1982, p. 14), the Police Academy uses experienced police officers as raters rather than psychologists or other private consultants. Nominated by their respective departments, officers are selected on the basis of experience, professional competence, and willingness to become involved in the assessment center. Once selected, they are required to attend the Police Academy and successfully complete a comprehensive five-day training course and a short apprenticeship with an experienced rater. Subsequently, as necessary, qualified raters are temporarily seconded to operational assessment centers, where, in addition to assessing candidates, they are required to participate as role players in the exercises. Notably, no raters are permitted to assess candidates with whom they are personally acquainted.

Candidates. Because the assessment center is an expensive and time consuming process, police departments only send those recruit candidates who have successfully

completed the first phase of the recruiting process. This initial phase generally includes the following components: (1) meeting minimum requirements with respect to age, education, citizenship, and driving record; (2) passing a criminal record check; (3) passing a battery of tests with respect to mental ability and/or academic ability, and language skills; (4) passing a physical abilities test; and (5) passing a primary interview. As a result of this screening process, range restriction is likely a factor at the assessment center (discussed in Chapter 3). If successful at the assessment center, candidates usually proceed to the final phase of the recruitment process, which generally includes a final panel interview, polygraph examination, medical examination, and security clearance investigation.

Structure

As discussed in Chapter 2, an assessment center is a standardized procedure, not a place. However, for convenience, operational assessment centers for police recruits are generally conducted at the Police Academy, where there are classrooms of various sizes, a simulation area that is permanently configured as a one-bedroom apartment (along with one-way viewing windows), and a cafeteria. The Police Academy, which is part of the Justice Institute of British Columbia (JIBC), is located in New Westminster, British Columbia, Canada.

An operational assessment center typically consists of approximately 10 to 12 candidates, six to eight raters, one or two administrators, and support staff. In such cases, the assessment center is usually divided into two groups, each composed of five or six candidates and three or four raters, ensuring that each candidate is assessed by multiple raters. A recruit assessment center requires two and one-half days. In the first day, the

raters observe and record the behavior of the candidates, where each rater observes each candidate in at least one exercise. In the second day, the raters prepare a final written report for each candidate, which includes classifying behavior by dimension and rating behavior according to a prescribed scale. And in the morning of the third day, the administrator and the raters discuss their findings in an integration session, where, for each candidate, they decide on a final score for each dimension and an overall assessment rating.

Based on the results of the integration session, the administrator prepares a final summary report for each candidate, which includes the original ratings given by each rater, the overall assessment rating, and a brief narrative. A copy of this report is sent to the participating police department, while the original is retained and filed by the Police Academy.

The Exercises

Although based on traditional exercises for assessment centers (see Chapter 2), Police Academy exercises are tailored to simulate critical activities that a police officer would encounter while on general patrol duties. The purpose of the exercises is to provide candidates with the opportunity to demonstrate behavior showing that they have the basic abilities necessary for success on the job. Various exercises are used to ensure that there is an adequate opportunity for the raters to observe all the dimensions of interest. The Police Academy uses five different exercises (leaderless group discussion, fact finding, oral communication, written report, and background interview), which are described in Table 5-1 below.

Table 5-1

Assessment center exercises

Exercises	Descriptions
1. Leaderless group discussion	Group exercise that requires the candidates to work together to develop a prioritized list of some particular policing objective.
2. Fact finding	Sometimes known as the "decision making" exercise, it requires individual candidates to gather and record factual information by interviewing a suspect in a minor investigation (e.g., theft complaint). Before this exercise, candidates are instructed in some basic, relevant legal principles, which must be applied in the exercise.
3. Oral communication	Similar to the "fact finding" exercise, except that it requires individual candidates to communicate some sensitive, often tragic, information to a citizen (e.g., death notification).
4. Written report	Requires individual candidates to write a report outlining their investigation in the "fact finding" exercise. This report, graded by either the administrator or a consultant, is used to assess the candidate's ability to synthesize relevant information and to write an accurate report, using good grammatical form.
5. Background interview	A structured behavioral interview that examines each candidate on biographical data so that additional information on the dimensions of interest can be added to the assessment. It is also typically used to probe the integrity of the candidate.

Originally, in 1977 there were six exercises, which included the leaderless group discussion, fact finding, observational, oral communication, interpersonal behavior, and background interview exercises. But in 1996 the fact finding and observational exercises were combined because they were repetitious, as were the interpersonal behavior and oral communication exercises. Also in 1996, written communication was introduced as an exercise. Because the Police Academy conducts a traditional, dimension-based assessment center, the raters do not assess the candidates on exercises but on dimensions and only after all exercises have been completed. Consequently, the change to exercises theoretically should not affect the analysis of reliability or discrimination.

The Dimensions

Consistent with traditional assessment center theory, the Police Academy defines a dimension as "a discreet, measurable portion of individual behavior which is associated

with success or failure on a job or at a job level" (Justice Institute of British Columbia Police Academy, 1997, p. 11). Developed through a "four-phased job analysis plan" that involved focus groups and surveys of police personnel in 1977, the Academy identified 15 critical dimensions (rf. Turner & Higgins, 1977), most of which are described in Table 5-2 (revised in 1994).

Table 5-2

Assessment center dimensions*

Dimensions	Descriptions
1. Ability to learn	Understands new information and applies it as necessary.
2. Decisiveness	Makes decisions, takes action and commits to it.
3. Fact finding/observation	Identifies and recalls relevant facts of an incident.
4. Flexibility	Adapts to new situations; modifies approach accordingly.
5. Initiative	Actively influences events versus passively acquiescing.
6. Integrity	Displays honesty and trustworthiness at all times.
7. Interpersonal sensitivity	Displays empathy and tolerance to all persons.
8. Maturity	Displays confidence, stability and sense of responsibility.
9. Oral communication	Expresses oneself and also listens attentively to others.
10. Personal impact	Projects good impression; commands attention, respect.
11. Practical intelligence	Analyzes problems; applies practical, logical solutions.
12. Problem confrontation	Addresses all situations, even if unpleasant or dangerous.
13. Stress tolerance	Maintains composure and performance under stress.
14. Written communication	Expresses oneself well in writing; good grammar.

* Revised in 1994.

The purpose of the dimensions is to enable raters to decide on an overall assessment score, which if valid will theoretically predict success as a police constable. Various dimensions are used so that all important aspects of the job (i.e., constable) are assessed (similar to multiple items or sub-scales in a test). It is not expected that candidates will necessarily display all dimensions in every exercise, but by incorporating different exercises raters are provided with a greater opportunity to assess the candidates on the dimensions of interest.

Originally in 1977 there were 15 dimensions, including the dimension of interpersonal tolerance. But because this dimension was apparently measuring the same

dimension as interpersonal sensitivity, both were combined in 1994, making a total of 14 dimensions. Also in 1994, the dimension known as adherence to authority (willingness to follow the rules inherent in a para-military organization) was replaced with maturity. Table 5-3 was adapted from Police Academy training material (Justice Institute of British Columbia Police Academy, 1997, pp. 6-12), which summarizes the dimensions most likely to be observed in each exercise.

Table 5-3

Dimension-exercise matrix

	Exercises				
Dimensions	GD	FF	OC	WE	BI
1. Ability to learn		x	x	x	x
2. Decisiveness	x	x	x		
3. Fact finding skills		x	x	x	
4. Flexibility	x	x	x		
5. Initiative		x	x		x
6. Integrity					x
7. Interpersonal sensitivity	x	x	x		x
8. Maturity		x	x		x
9. Oral Communication	x	x	x		x
10. Personal impact	x	x	x		x
11. Practical intelligence	x	x	x	x	x
12. Problem confrontation		x	x		
13. Stress tolerance	x	x	x		x
14. Written communication				x	x

Note: x indicates the dimension that will likely be observed.

Legend:

GD: Leaderless group discussion
FF: Fact finding exercise
OC: Oral communication exercise
WE: Written exercise
BI: Background interview

The Scoring Procedure

For over 20 years the Police Academy assessment center has not departed from the traditional dimension-based approach.[2] In this approach, the intent of the exercises is to provide the opportunity for candidates to display behavior that the raters classify

[2] See Gavin and Hamilton (1975, p. 172) for an example of the dimensions used in a traditional dimension-based police assessment center in 1975 in Colorado. Interestingly, the dimensions are almost identical.

according to the pre-defined dimensions. The raters make copious notes during an exercise, but training and management emphasize the classification of behavior according to dimensions rather than exercises. This procedure is reinforced by the practice of suspending assessments until all documentation on a candidate has been completed.

After all candidates have completed the exercises, the raters, are given one full day to organize their notes, review their behavioral classifications, and subsequently assign a numerical rating for each dimension. These dimension ratings are not norm-referenced nor are they comparative within a class. Rather, raters assess behavior against pre-defined standards for each dimension (i.e., the assessment is criterion-referenced). Notably, the definitions provided in Table 5-2 are only summaries—each rater is provided with a set of comprehensive performance standards that suggests a rating based on the level displayed by a candidate. The dimensions are not weighted and collectively are assumed to be measuring the potential for success as a police constable.

Finally, in the integration session (which is moderated by the assessment center administrator), the raters review their notes for each candidate. These notes are an important record of candidate behavior and provide the rationale for the rater assessments. During the integration session, any rater, based on the adequacy of the record, may challenge another rater's assessments. After the raters agree on a final rating for each dimension for each candidate, the raters decide on an overall assessment rating for each candidate. The overall assessment rating is not an arithmetic average of the dimensions (although it is usually similar); rather, based on a compendium of evidence, it reflects the raters' best judgment on whether the candidate will be successful as a police constable.

Although the administrator prepares a written summary report for each candidate, because most participating police departments use an overall assessment rating of 38 (slightly above average) as a cut score, in reality a candidate's progress in the recruitment process usually depends on this score alone (see Table 5-4 below, which reproduces Table 3-1 from Chapter 3).

Table 5-4

Police Academy assessment center rating scale

Rating	Score	Converted Score*
Excellent	5	55
Superior ability	5-	52
A great deal of ability	4+	48
Well above average[+]	4	45
Above average	4-	42
Slightly above average	3+	38
Average (competent)	3	35
Slightly below average	3-	32
Below average	2+	28
Well below average	2	25
Very little ability	2-	22
Poor	1+	18
Very poor	1	15

* Rather than converting base scores to decimals (e.g., 3+ to 3.33), the Police Academy uses the "converted score."

[+] The term "average" should be read as "competent."

Table 5-4 indicates a thirteen-point rating scale, but as discussed in Chapter 3 this scale is more realistically defined in terms of eleven points. Over 22 years, 99.6% of assessments were limited to nine points (22 to 48) and 97% were limited to seven points (22 to 42).[3] Apart from explanations such as central tendency and range restriction, Sergeant MacDonald, the current assessment center administrator, provided the

[3] A rating scale of seven to nine points is more consistent with other assessment centers. For example, Ross (1980) reports a seven-point rating scale in her study (e.g., unacceptable, marginal, fair, average, good, excellent, and superior), as does Friedman (1984), while Huck and Bray (1976) report seven- to nine-point rating scales.

explanation that the lowest ratings are rarely given, even if deserved, because raters "do not want to destroy someone," while the highest ratings are unattainable (personal communication, November, 1999).

With the assessment center explained (Chapter 2), reliability and discrimination defined (Chapters 3 and 4), and the context here provided, it is now possible to specify the research questions as they relate to the particular objectives of assessing the Police Academy on the matters of interrater reliability and systemic sex discrimination.

Assessment of Interrater Reliability

The first problem addresses the issue of whether the Police Academy assessment center meets the conventionally accepted standard of .60 for adequate overall interrater reliability. Specifically, the question is whether there is a high degree of relative consistency (i.e., correlation) amongst raters when assessing candidates. Based on the literature (rf. Chapter 3), it is hypothesized that the corrected reliability (R) is equal to or greater than .60 (i.e., $R \geq .60$). The second problem, which addresses the issue of discrimination, will be discussed later in this chapter.

Sample and Data collection

The sample consists of 55 operational assessment centers (ACs)[4] that were conducted between 1978[5] and October of 1999. These ACs were designed exclusively to screen candidates who were competing for entry-level positions in municipal police

[4] Generally, the term assessment center refers to a personnel selection method. However, by convention this term also applies to what has previously been described as an operational assessment center, in which a particular group of candidates are assessed. The meaning is usually clear from the context, but for the sake of brevity and clarity, from this point on an operational assessment center will be designated as "AC," while the term assessment center will be used to reflect its more general meaning.

[5] The assessment center was organized in 1977; however, the first group was not processed until 1978.

departments in British Columbia. As shown in Tables 5-5 and 5-6 (which provide data summaries for candidates and raters), on average, three or four raters assessed five or six candidates in each AC.

Table 5-5

ACs (1978 to 1999): Candidate summaries

	N (ACs)	Min.	Max.	Mean	Total
Male	55	1	8	4.27	235
Female	55	0	5	1.00	55
Totals	55	3	8	5.27	290

Table 5-6

ACs (1978 to 1999): Rater summaries

	N (ACs)	Min.	Max.	Mean	Total
Raters	55	2	6	3.80	209*

* Because raters are drawn from a limited pool, the same rater may be counted more than once.

The Police Academy does not store AC results electronically, except for candidate identity and AC number. Rather, all information is recorded on paper, classified by AC number, and filed in locked cabinets. After one or two years, the files are transferred to cardboard boxes, sealed, and stored in government warehouses. In order to conduct this study, arrangements were made with the Police Academy director and assessment center administrator to access current records and retrieve archived records. Subsequently, summary sheets for the sample were located, copied, and the data subsequently entered into SPSS® Base 10.0.

Systematic random sampling was used to select the sample of ACs. Specifically, every fifth AC was selected, beginning with AC #5 in 1978 and ending with AC #262 in 1999, for a total of 55. Apparent inconsistencies in increments (e.g., ending with AC #262) are the result of either numbering irregularities by the Police Academy or split assessment centers. AC #10 was omitted because data were missing. Sometimes the administrator divided an AC into groups so that a set of raters would not have to assess more than eight candidates (e.g., if there were 12 candidates the AC would typically be divided into two classes of six candidates each). If this were the case, for sampling purposes one group was selected by means of a random number table. Any group that consisted of less than three candidates was not considered eligible for selection.

When written communication was introduced as an exercise in 1996, Police Academy practice was that a separate rater scored only the dimension of written communications for this exercise. In these cases, that rater was not included in the analysis as it could artificially inflate the corrected (effective) reliability coefficient.

Data Analysis

In this study, interrater reliability is defined as the consistency of overall assessments made by individual raters, within a fixed set, on candidates in a specific AC group. At the Police Academy, individual raters do not technically provide overall scores for each candidate. As a result, to determine a rater's overall score for a candidate it was necessary to calculate the mean of the dimension scores that a rater gave for each candidate. Notably, reliability coefficients are not artificially inflated because they were calculated from rater assessments that were made independent of any consensus reached in an integration session.

Interrater reliability coefficients for each dimension were not calculated for two reasons. First, because raters do not observe each candidate in all five exercises (usually only one or two) they do not have the opportunity to score all specified dimensions for each candidate. Second, raters exercise their own discretion in deciding whether or not there is sufficient evidence to score a particular dimension. As a result, dimensions were often scored by only one rater, which made it impracticable to calculate the reliabilities of individual dimensions in any systematic way.

Using the reliability program in SPSS® Base 10.0, interrater reliability coefficients for each AC were calculated by means of the following three techniques.

1) Average correlation (i.e., the "correlation approach"): corrected (Spearman-Brown formula) reliability R for a set of raters.
2) Cronbach's alpha (α): "true alpha" (as opposed to "standardized item alpha," which simplifies to R above).
3) Intraclass: both $\bar{r}$ ($ICC_{3,1}$), the mean reliability of a single rater, and R ($ICC_{3,k}$), the corrected reliability for a set of raters. Because the raters were "fixed" (they are the only raters of interest), intraclass configuration #3 using a "mixed effects" model was used (rf. Chapter 3).[6]

In addition, based on all ACs collectively in the sample (vis-a-vis separate ACs), configuration #1 of the intraclass technique was used to determine the mean reliability $\bar{r}$ of a single rater ($ICC_{1,1}$), and the corrected reliability R for a set of raters ($ICC_{1,k}$) (rf. Chapter 3). Because this configuration uses a random effects model (one-way ANOVA),

[6] As noted in Chapter 3, SPSS® Base 10.0 uses two-way ANOVA.

it is possible to generalize the results to the population of Academy raters because it assumes that raters were selected randomly from the population.

Police Academy AC reliability was assessed by comparing these coefficients with the conventionally accepted standard of .60 (for corrected overall reliability). The mean reliabilities for each year and the mean reliability for all combined years were calculated for the purpose of making specific and general comparative assessments. Because reliability coefficients are estimates of the correlation between measurements obtained on a target (Fleenor, Fleenor & Grossnickle, 1996, pp. 373-374; Rosenthal & Rosnow, 1991, pp. 50, 431; Shrout & Fleiss, 1979, p. 422; and Tinsley & Weiss, 1975, p. 363; see also Rosnow & Rosenthal, 1996, p. 126), mean reliabilities were calculated using Fisher's *Z* transformations (rf. Greenwood & McNamara, 1967; and Hinrichs & Haanpera, 1976). Such transformations are not always necessary, but because the differences between many coefficients were large and because many of the individual coefficients were large (over .50), a simple arithmetic mean would have grossly underestimated reliability.

Because three different techniques were used, alpha and intraclass *R* ($ICC_{3,k}$) were compared with the correlational approach *R* to assess its utility as an uncomplicated yet appropriate technique to estimate interrater reliability. As discussed in Chapter 3, there are four general assumptions associated with parametric tests of reliability (i.e., all items are assumed to be equivalent, all items are assumed to be measuring the same dimension, all errors are assumed to be random and unrelated, and an item score is assumed to be the sum of its own true score and error score). The first three assumptions are likely met approximately, but the last is strictly theoretical and so the degree to which it is met cannot be known.

In addition to assessing interrater reliability, based on the collective sample the following two analyses, as suggested by Hinrichs and Haanpera (1976), were conducted on the scale itself for the purpose of determining if improvements could be made:

1) item (dimension) analysis, where bivariate correlations (r) between the final ratings for each dimension and the overall assessment ratings (OAR)[7] were computed; and
2) internal (scale) consistency analysis, where average bivariate correlations ($\bar{r}$) between dimensions were computed, summarized by Cronbach's alpha (true and standardized item).

Correlational item analysis was possible without adjusting for artificially inflated coefficients because the dimension scores and OARs were mathematically independent of each other. Specifically, the overall rating is not a quantitative average of the dimension scores, but represents a consensus among raters reached during the integration session and so is mathematically independent from the dimension scores (rf. Hinrichs & Haanpera, 1976, p. 37).

Assessment of Discrimination

The second problem addresses the issue of whether the Police Academy assessment center discriminates on the prohibited ground of sex, and so is appropriate for correlational analysis to determine if there is a systematic relationship between assessment center performance and sex. Specifically, the question is whether OAR is a function of candidate sex. Based on the literature (rf. Chapter 2), it is hypothesized that

[7] For brevity and clarity, overall assessment rating will generally be referred to as OAR.

there is a correlation between sex and OAR; specifically, that higher OARs will be correlated to females (i.e., $\rho_{xy} \neq 0$). However, it is further hypothesized that when mediating (i.e., intervening) variables such as age, education, and cognitive ability are controlled, the correlation between sex and OAR will be insignificant (i.e., $\rho_{xy \cdot z} = 0$).

Sample and Data Collection

In order to assess sex discrimination, access was gained to the records of the entire population of 2,956 police recruit candidates who attended Police Academy ACs from the first class in 1978 to October of 1999. Notably, candidates cannot nominate themselves to attend, but must be sponsored by one of the 12 municipal police departments in British Columbia. Department size, measured by total number of sworn police officers, ranges from approximately 20 to 75 (Central Saanich, Esquimalt, Nelson, Oak Bay, Port Moody, West Vancouver), 100 to 175 (Abbostford, Delta, New Westminster, Saanich, Victoria) and 1,200 (Vancouver). There is now a relatively small First Nations municipal police department (Stl'Atl'Imx); however, at the time of this study it was not yet designated as an independent police department by the Ministry of the Attorney General. There are other First Nations public safety services, but these operate under special memoranda of agreement with the Ministry of the Attorney General.

Because the Police Academy does not store records electronically (previously discussed), it was necessary to physically search paper files to collect the necessary data, which were subsequently entered into SPSS® Base 10.0. Upon searching these files it was found that information on a candidate was limited to name, date of assessment, dimension scores, OAR, sponsoring department, and a self-developing (Polaroid)

photograph. Therefore, to conduct the study, it was necessary to obtain information regarding a candidate's sex, age, years of post secondary education, and Police Intake Exam score from other sources.[8] Not unexpectedly, this information was limited because of incomplete records, unavailability of data, and logistical considerations. Notwithstanding, as discussed below, sufficient information was obtained to conduct the necessary analyses on these variables.

Because AC files included first and last names, along with photographs, it was possible to determine sex in all but five cases by comparing candidate names with corresponding photographs. As indicated in Table 5-7 (which shows male and female attendance in three-year intervals), of the total of 2,956 candidates, 2,478 (83.8%) were determined to be male and 473 (16%) were determined to be female, with data missing on 5 (0.2%) cases. Notably, the rate of attendance by females at ACs has increased steadily over time, from 8.9% in 1978-80 to a high of 24% in 1996-98, although it has leveled off to approximately 20%. The explanation for this trend is complex (involving variables such as social values, interest by women, police recruitment objectives, and laws against discrimination) and beyond the scope of this study (e.g., see Landrine & Klonoff, 1997, p. ix).[9] In any event, there has been an increase in applications by women and municipal police departments have responded by sending more women to the assessment center. According to anecdotal information by police personnel officers, the

[8] Reliable information about a candidate's race was not available.

[9] When analyzing personnel selection decisions for discrimination, it is generally accepted that the existing pool of applicants is the appropriate population for two reasons: first, because it provides direct evidence about the employer's selection process; and second, because to do otherwise would "penalize employers for social phenomena somewhat external to them" (Paetzold & Willborn, 1999, §§ 4.03, 5.04; rf. also Baldus & Cole, 1980; for a Canadian perspective, see Vizkelety, 1987, pp. 184-185, comparing Action Travail, 1987, with Blake, 1984).

proportion of women sent to the assessment center corresponds to the proportion of women who apply for entry level positions, which appears to have leveled off since 1993 (between 21% and 25%).[10]

Table 5-7

Candidate statistics (1978-1999): Totals for males and females

Year	Males		Females		Total		Missing	
	N	%	N	%	N	%	N	%
1978-80	270	91.10	27	8.90	297	100	0	0
1981-83	245	93.87	16	6.13	261	100	0	0
1984-86	223	92.53	18	7.47	241	100	0	0
1987-89	413	88.25	55	11.75	468	100	0	0
1990-92	522	82.73	109	17.27	631	100	0	0
1993-95	277	76.31	86	23.69	363	100	0	0
1996-98	405	75.00	130	24.00	535	99.0	5	1.00
1999*	123	79.35	32	20.65	155	100	0	0
Totals	2478	83.80	473	16.00	2951	99.80	5	0.20

* Up to October, 1999.

Because municipal police departments in British Columbia are required by law to send all police recruits to the Police Academy for basic training, information on candidate age and years of post secondary education was obtained by searching Police Academy training records. As a result, this information was generally limited to those candidates who were hired by a municipal police department in British Columbia.

[10] A 20% recruitment rate for females is consistent with that reported by Dantzker and Kubin (1998) in the Unite States (p. 19).

As indicated in Table 5-8 (which shows candidate age for males and females in three-year intervals), candidate age was determined in 1,166 cases (39.45%) out of a total of 2,956 cases, with males accounting for 952 (32.21%) cases and females accounting for 214 (7.24%) cases. Of interest is that candidate age has steadily increased over time, from an average age of 24.98 in 1978-80 to an average age of 28.03 in 1996-98. Notably, female candidates generally have been slightly older than male candidates, the overall average age of females and males being 26.49 and 25.61 respectively.

Table 5-8

Candidate statistics (1978-1999): Age, grouped by sex

	Male			Female			Totals		
Year	Mn	Sd	Valid N	Mn	Sd	Valid N	Mn	Sd	Valid N
1978-80	24.95	3.95	154	25.45	2.88	11	24.98	3.88	165
1981-83	24.31	3.66	149	22.70	2.79	10	24.21	3.63	159
1984-86	25.53	3.39	127	26.00	4.33	9	25.56	3.44	136
1987-89	25.77	3.97	176	25.08	3.14	38	25.64	3.84	214
1990-92	25.57	4.01	193	26.73	4.09	48	25.80	4.04	241
1993-95	26.96	4.56	70	26.82	3.56	50	26.90	4.16	120
1996-98	27.98	4.16	83	28.13	3.75	48	28.03	4.00	131
1999	.	.	.	.	.	.	.	.	.
Totals	25.61	4.03	952	26.49	3.82	214	25.77	4.00	1166

Note: . indicates no data available.

As indicated in Table 5-9 (which shows candidate education for males and females in three-year intervals), level of candidate education (defined as years of attendance at an accredited post secondary institution) was determined in 1,104 cases (37.34%) out of a total of 2,956 cases, with males accounting for 898 (30.37%) cases and

females accounting for 206 (6.97%) cases. Similar to the pattern found for age, candidate education has steadily increased over time, from an average of 1.69 years in 1978-80 to an average of 3.81 years in 1996-98. Of particular interest is that female candidates, on average, have been slightly better educated than male candidates, the overall average years of post secondary education of females and males being 3.24 and 2.50 respectively.

Table 5-9

Candidate statistics (1978-1999): University education (years), grouped by sex

	Male			Female			Totals		
Year	Mn	Sd	Valid N	Mn	Sd	Valid N	Mn	Sd	Valid N
1978-80	1.60	1.71	119	3.00	1.77	8	1.69	1.74	127
1981-83	1.47	1.34	148	2.22	1.30	9	1.52	1.34	157
1984-86	2.40	1.59	122	2.63	1.60	8	2.42	1.59	130
1987-89	2.62	1.51	168	2.53	1.56	36	2.60	1.51	204
1990-92	2.87	1.44	188	3.10	1.57	48	2.92	1.46	236
1993-95	3.67	1.34	70	3.46	1.47	48	3.58	1.39	118
1996-98	3.69	1.35	83	4.02	1.09	49	3.81	1.27	132
1999	.	.	.	.	.	.	.	.	.
Totals	2.50	1.65	898	3.24	1.53	206	2.64	1.66	1104

Note: . indicates no data available.

With respect to Police Intake Exam scores, information was even more limited. Although the Police Academy constructed this exam and controlled its distribution, it did not maintain a user data file. To obtain a sample of candidates it was necessary to obtain permission from the Vancouver Police Department to access their recruiting records. Because Vancouver Police is the largest municipal police department in the province, with approximately 1,200 sworn members, it was the logical choice to obtain a suitable

sample for analysis. Here, every 20th applicant from 1978 to 1996 was selected. From this sample, every candidate who had attended an AC and was hired was selected. This resulted in a systematic random sample of 219 candidates, where the proportion of males and females was similar to the age and education samples (for case and data summaries, see Appendix 1 and Appendix 2). In other words, the Police Intake Exam sample was a random Vancouver sample (N (listwise) = 205) of the age and education non-random samples.

As indicated in Table 5-10 (which shows candidate scores on the Police Intake Exam for males and females in three-year intervals), of the 219 candidates (7.41% of the total sample of 2,956 candidates), males accounted for 171 (5.79%) cases and females accounted for 48 (1.62%) cases.

Table 5-10

Candidate statistics (1978-1999): Police Intake Exam scores (%), grouped by sex

Year	Male			Female			Totals		
	Mn	Sd	Valid N	Mn	Sd	Valid N	Mn	Sd	Valid N
1978-80	66.13	6.17	15	68.00	0.00	1	66.25	5.98	16
1981-83	66.58	9.05	19	64.33	1.53	3	66.27	8.43	22
1984-86	72.23	9.68	13	.	.	.	72.23	9.68	13
1987-89	68.13	8.75	23	70.80	4.49	5	68.61	8.15	28
1990-92	71.35	7.59	86	74.00	8.58	21	71.87	7.82	107
1993-95	72.50	12.39	10	75.59	7.36	17	74.44	9.42	27
1996-98	75.80	9.42	5	73.00	0.00	1	75.33	8.50	6
1999	.	.	.	.	.	.	.	.	.
Totals	70.19	8.55	171	73.48	7.75	48	70.91	8.47	219

Note: . indicates no data available.

Consistent with the pattern found for candidate education, Police Intake Exam scores have increased over time, from an average score of 66.25 in 1978-80 to an average score of 75.33 in 1996-98. Again, of particular interest is that female candidates, on average, scored somewhat higher than male candidates, the overall average scores (reported by way of percent) of females and males being 73.48 and 70.19, respectively.

Police Intake Exam scores were important in an attempt to control for mental ability/academic skills when exploring the relationship between sex and OAR. In the mid-1970s the Police Academy constructed the Police Intake Exam, which from 1978 to 1998 was used extensively by municipal police departments as a screening instrument.[11] This test was originally named the Police Intake Level Examination but was commonly referred to as the Police Intake Exam. It was also sometimes referred to as the "Police Educational Intake Examination," which provided a somewhat more accurate description as it resembled the Canadian "General Educational Development" (GED) test for high school equivalency (e.g., see Rockowitz, Shuttleworth, Shukyn, Brownstein & Peters, 1998).

For example, the GED tests in five general areas, including English usage and writing skills, social studies, science, literature and the arts, and mathematics. The Police Intake Exam tests for English usage and writing skills, logic analysis, comprehension and reasoning, memory, and mathematics. According to Rockowitz et al. (1998), rather than testing absolute recall of facts, the purpose of the GED is to measure one's ability to understand and apply information, evaluate, analyze, and draw conclusions, and express

[11] In 1998, the Police Academy and some police departments began using a recently introduced Canadian version of the Wonderlic Personnel Test (WPT) (1992), which is a mental ability test. However, some police departments continue to use the Police Intake Exam.

ideas and opinions in writing (p. 6). Similarly, the purpose of the Police Intake Exam is to measure logical reasoning, practical judgment, and level of academic skills (Mr. K. Taylor (a Police Academy employee who was primarily responsible for the construction of the Police Intake Exam), personal communication, July, 1981).

Coincidentally, mental ability tests such as intelligence tests and general aptitude tests claim to measure practical or general intelligence, a factor that theoretically represents thinking capacity, such as reasoning and problem solving ability, but which is difficult to disentangle from academic aptitude (Kachigan, 1991, p. 158; Singer, 1993, p. 22).[12] It seems reasonable, then, to assume on its face that the Police Intake Exam (and the GED), in addition to academic skills, likely measures to some degree a candidate's level of mental ability. As noted above, the Police Academy did not maintain a user data file for the Police Intake Exam and kept almost no documentation of its historical development. As a result, there is no evidence of its reliability, validity, or discriminatory effect, nor is it norm-referenced.

These limitations notwithstanding, the Police Intake Exam was the only measure available to control for mental ability, whether it is defined as intelligence or academic skills or some combination of both. In defense of the Police Academy, it has never made any theoretical claims regarding the exam's ability to predict mental ability, although it was implied that it could predict academic ability (arguably for the purpose of predicting success in recruit training at the Police Academy). In any case, the Police Academy was

[12] As was noted in Chapter 4, it has also been argued that mental ability tests are actually tests of culturally based educational skills (e.g., middle class educational values) (rf. Singer, 1993, p. 23, who reviews Eysenck, 1984; Jensen, 1985; and Wigdor & Garner, 1982). This topic, however, is beyond the scope of this study.

apparently caught unprepared by the popularity of the Police Intake Exam, which continues to be used as a screening instrument in the municipal police community.

In summary, the samples for age and education each represented over 37% of the population, while the sample for the Police Intake Exam represented only 7.41% of the population (see Appendix 1). The candidates in the age and education samples were not randomly selected, but were those who attended the Police Academy assessment center and were subsequently hired by a municipal police department. Similarly, the sample for the Police Intake Exam, although of sufficient size, was not random—it was a systematic random subset (Vancouver Police Department candidates) of a nonrandom set (all candidates who had attended the Police Academy assessment center).

However, because population data were available for OAR and sex, it was possible to compare the various samples with the population on these two variables (see Table 5-11, which, for males and females, shows OAR summaries for each sample and the population). Not unexpectedly, the total mean OAR of each sample, representing those candidates who were successful in the hiring process, was higher than the total mean OAR of the population. Nevertheless, the pattern of mean OARs in each sample was similar to that found in the population, where mean OARs of females were consistently higher than mean OARs of males. Moreover, the distribution of males and females in each sample was similar to that found in the population, where the proportion of male and female attendance was approximately 80% and 20% respectively. For these

reasons, and in the absence of evidence to the contrary, the samples appeared suitable for the purposes of this study.[13]

Table 5-11

Comparison of sample to population on OAR, grouped by sex

	N	%	Min	Max	Mn	Sd
Population[a]						
Male	2468	83.9	15	48	36.55	4.31
Female	473	16.1	25	48	37.95	3.79
Total	2941	100	15	48	36.78	4.26
Excluded (Age & Educ.)						
Male	1514	85.4	15	48	35.52	4.48
Female	258	14.6	25	48	36.93	3.96
Total	1772	100	15	48	35.73	4.44
Not Hired						
Male	1114	87.5	15	48	34.98	4.67
Female	159	12.5	25	48	36.96	4.23
Total	1273	100	15	48	35.23	4.66
Sample (Age)[b]						
Male	952	81.7	22	48	38.17	3.44
Female	214	18.3	28	45	39.16	3.19
Total	1166	100	22	48	38.35	3.42
Sample (Education)[b]						
Male	898	81.3	25	48	38.25	3.33
Female	206	18.7	28	45	39.19	3.19
Total	1104	100	25	48	38.42	3.32
Sample (Age & Educ)[c]						
Male	896	81.4	25	48	38.24	3.32
Female	205	18.6	28	45	39.17	3.19
Total	1101	100	25	48	38.41	3.32
Sample (Exam)[d]						
Male	171	78.1	28	45	37.89	3.13
Female	48	21.9	28	45	38.62	3.84
Total	219	100	28	45	38.05	3.30
Sample (Age, Educ & Exam)[e]						
Male	159	77.6	28	45	37.92	3.04
Female	46	22.4	28	45	38.41	3.76
Total	205	100	28	45	38.03	3.21

Notes:
[a] All candidates (1978-99) who attended the Police Academy recruit assessment center.
[b] Candidates who attended the Police Academy, usually for recruit training.
[c] Listwise sample, for which data existed for both age and education.
[d] Systematic random sample of Vancouver Police recruit files, and is a subset of the age and education samples.
[e] Listwise sample, for which data existed for age, education, and Police Intake Exam.

[13] Paetzold and Willborn (1999), also citing Baldus and Cole (1980), argue in discrimination cases that "unless there is evidence that the 'sample' is not representative, it is acceptable to view statistical evidence from the sample as representing an inference about the relevant population" (§2.05, footnote 36).

Data Analysis

Systemic discrimination, which arises at the aggregate level, may be operationally defined as an unjustified discriminatory effect against a protected group; that is, on prohibited grounds there exists prima facie evidence of prejudicial distinctions (whether overt or subtle, intended or unintended) between groups (e.g., males and females). Given that discrimination is a comparative concept where the variables of interest can often be measured on a numerical scale, a quantitative analysis may provide material evidence in finding for or against discrimination. Generally, such an analysis ranges from calculating simple differences to tests of statistical significance.[14] Specifically, with respect to this study, discrimination was quantitatively assessed by means of a parity analysis and a probity analysis,[15] which are described in detail below.

Parity (Comparative Evidence)

First, the mean OARs (1978-88; 1989-99; 1978-99) of female candidates were compared to those of male candidates to determine if systematic differences existed between them. Second, using a cross tabulation to summarize the data (1978-99), the

---

[14] This study reflects the traditional approach to a statistical analysis of discrimination, which in turn is based on traditional notions of probability. There is, however, a non-traditional school of thought known as Bayesian analysis (rf. Lee, 1997), which has recently been debated by legal commentators and statisticians (Vining, McPhillips & Boardman, 1986, pp. 687-688), but which has received little attention in reported cases (Paetzold & Willborn, 1999, §§ 12.01-12.05). The Bayesian approach to probability (i.e., prior probabilities) accounts for an individual's degree of belief, as it acknowledges that factors other than frequency are grounds for making inferences. As a result, within the context of discrimination, the Bayesian approach allows for direct rather than indirect inferences, as is the case in the traditional approach. Disadvantages of the Bayesian approach include difficulties in determining "prior probabilities," computational complexities, and lack of consensus on its utility (e.g., see Baldus & Cole, 1980).

[15] These steps were introduced in Chapter 4, in the section "Assessing Systemic Sex Discrimination." Notably, probity is not conventionally a legal or statistical term, but in this study refers to a test of the integrity of the selection process.

OAR distribution (by category) of female candidates was compared to that of male candidates to determine if systematic differences existed between them.[16]

Probity (Legal/Practical Significance)

This step included tests that analyzed whether or not any differences found in the parity analysis were legally significant. The probity analysis included (1) the four-fifths rule (which uses an absolute standard), (2) statistical tests (which use alpha levels), and (3) the binomial effect-size display (which translates *r* into practical significance).

1. The four-fifths rule. The four-fifths rule was calculated using the procedure described below (rf. Uniform Guidelines, 1978).[17]

1) Pass rates were calculated for each sub-group within the pool of candidates. For example, if 40 male candidates out of 100 male candidates and 15 female candidates out of 50 female candidates passed (the total candidate pool being 150), the pass rates would be .40 and .30 respectively.
2) The dominant group was identified, which in the example above would be the male candidates who had a pass rate of .40.
3) The adverse effect rate was calculated by dividing the pass rate of a sub-group by the pass rate of the dominant sub-group (e.g., .30 / .40 = .75).
4) The sub-group, whose pass rate was below .80 of the dominant group, was identified. For example, the female candidate pass rate of .30 was below .80 of the male candidate pass rate of .40; therefore, according to the four-fifths

[16] Before the two-step procedure was formalized by U.S. courts, researchers such as Huck and Bray (1976), Jaffee et al. (1972) and Moses and Boehm (1975) were using a comparable procedure to assess discrimination in assessment centers.

[17] For examples and commentary, see Hoffman and Thornton (1997), Hurley (1987), Magaldi, Mendoza, Stafford and Frank (1984), Paetzold and Willborn (1999, § 5.06), and Pynes and Bernardin (1992).

rule prima facie evidence of adverse effect against females was demonstrated. This general rule applies regardless of the membership of the sub-group (i.e., whether male or female).

The four-fifths procedure described above was applied to the OAR distribution (rf. Table 5-4), which was split at the median (38) to categorize it into dichotomous outcomes of pass and fail (which is generally how police managers interpret the results). Using a cross tabulation to summarize the data (1978-99), the selection rates were calculated for males and females, the dominant group identified, and the adverse effect rate determined.

2. Statistical tests. Because this study considers statistics within a legal framework, the following definitions should be noted. First, the term "statistics" refers to that "branch of mathematics dealing with the collection, presentation, and analysis of quantitative information" (rf. Paetzold & Willborn, 1999, § 2.01). Second, the term "statistical evidence" may refer either to simple comparisons or more complex tests. Third, the term "statistical tests" refers to those techniques used to make inferences (most of which were calculated by using the SPSS® Base 10.0 computer program).

Pearson's product-moment correlation *r*. First, to determine if a relationship existed between OAR and sex (i.e., whether OAR, the criterion variable, is a function of candidate sex, the predictor variable),[18] a bivariate correlation coefficient (r) between these two variables was calculated (e.g., see Parker, 1980).[19] In addition, exploratory

[18] Because this study is correlational rather than experimental, the variables are described in terms of predictor and criterion rather than in terms of dependent and independent.
[19] Because Pearson's r is the average value of products of paired z scores, the relationship between dissimilar variables measured on different scales can be assessed (Kachigan, 1992, p. 130).

bivariate correlations (pairwise) were calculated between OAR, sex, age, education, and Police Intake Exam score and presented in correlation matrices. Of particular interest, of course, were correlations that might explain differences between males and females on the OAR criterion. Second, the correlation between OAR and sex was tested for statistical significance (one-tailed), as were the remaining correlations (two-tailed) (rf. Miller & Whitehead, 1996, p. 325; and Erickson & Nosanchuk, 1977, p. 343). For each test of significance, where possible exact *p* values were reported, along with Bonferroni adjustments to protect against Type I errors (Rosenthal & Rosnow, 1991, pp. 329-332).[20] And third, for each bivariate correlation, *r* was squared to interpret the strength of the relationship (Green, Salkind & Akey, 1997, p. 283; Miller & Whitehead, 1996, p. 325; Rosnow & Rosenthal, 1996, pp. 259-260). Specifically, the r^2 coefficient (also known as the coefficient of determination) provided an estimate of the proportion of variance in the criterion (OAR) that was explained by each of the predictors (sex, age, education, and Police Intake Exam score).

The variables OAR, age, education, and Police Intake Exam score are continuous (at least at the quasi-interval level), while the variable sex is naturally dichotomous. As a result, point-biserial correlations were calculated when sex was correlated with another variable (Kachigan, 1991, pp. 189-190; Miller & Whitehead, 1996, p. 340; Paetzold &

---

[20] The use of statistical significance to assess discrimination was introduced in 1974 by the Equal Employment Opportunity Commission Guidelines, and the U.S. Courts typically accept .05 or less as statistically significant since it was first adopted by the U.S. Supreme Court in 1975 in Albermarle Paper Co. v. Moody (Paetzold & Willborn, 1999, §§ 4.11, 6.07; Vining et al., 1989, pp. 680-681). For example, in Castaneda v. Partida (1977) and Hazelwood School Dist. v. United States (1977) the threshold for statistical significance was identified as two or three standard deviations between the observed and the expected, which the U.S. Supreme Court described as a "gross statistical" disparity. In the U.S., Castaneda is often credited with approving the use of statistical analysis in discrimination litigation (Brook, 1990, p. 132); and although there is no comparable case by the Canadian Supreme Court, in Public Service (1999) the Court ruled that a "disproportionately negative effect" was sufficient for a finding of discrimination.

Willborn, 1999, § 6.09; Rosnow & Rosenthal, 1996, p. 242). Here, sex was quantified by arbitrarily coding males as 1 and females as 2, otherwise known as "dummy coding."

To the extent that one or more of the measurements on the variables of interest are unreliable, correlation coefficients will underestimate the true correlation (Borg & Gall, 1983, p. 593). More specifically, the correlation coefficient cannot exceed the square root of the product of the reliability coefficients of the variables of interest. Because OAR depends upon interrater reliability, which was estimated in this study, correlation coefficients calculated from OARs were corrected for attenuation. With the assumption that measurement error was random, correcting correlations for unreliability was calculated by the method outlined by Carmines and Zeller (1979) (rf. Chapter 3).

Tests of significance associated with a Pearson correlation coefficient between two variables make the following four assumptions: (1) variables are measured at an interval level; (2) variables are normally distributed (generally assumed for Pearson's *r*); (3) variables are bivariately normally distributed (i.e., otherwise known as a joint normal distribution, where each variable is normally distributed at all levels of the other variable, which results in a linear relationship); and (4) variables are random (i.e., they covary naturally and are not experimentally manipulated), where the scores on one variable are independent from those of the other (Erickson & Nosanchuk, 1977, pp. 241, 340; Green et al., 1997, p. 282; Kachigan, 1991, pp. 127-128; Koosis, 1997, p. 196; Miller & Whitehead, 1996, p. 329).[21]

[21] Inferential statistics, of course, assume that the sample was selected randomly.

All assumptions are generally met and have been previously discussed.[22] Except for sex (which was dummy-coded) the data are interval or at least quasi-interval, the sampling distribution of the mean will tend to be normal because of large sample sizes, and the variables have not been manipulated. One assumption requiring additional attention, however, is that of linearity, which was tested for significant deviation from linearity by means of one-way ANOVA (rf. SPSS® Base 10.0, 1999, pp. 100-101).

Here, a significant *p* value (.05) indicates that an hypothesis that group means are located on a straight line should be rejected, while a non-significant *p* value indicates that a linear relationship may be assumed. The *p* values in Table 5-12 (which provides the results of the ANOVA test for linearity) indicate that linear relationships exist between OAR and sex, OAR and Police Intake Exam, sex and age, sex and education, sex and Police Intake Exam, age and Police Intake Exam, and education and Police Intake Exam.

Table 5-12

Deviation from linearity: *p* values

	Sex	Age	Educ	Exam
OAR	.480	.001	.006	.217
Sex		.808	.719	.968
Age			.000	.449
Educ				.748

Note: $p < .05$ (without Bonferroni adjustments) indicates that no linear relationship exists.

[22] Courts in the U.S. have been willing to accept inferences about discrimination without ruling on the issue of underlying assumptions (Paetzold & Willborn, 1999, §6.04). As a result, Paetzold and Willborn argue that statistical analyses that are deficient in meeting assumptions still have utility in discrimination cases because the probative value of the analyses is a question of law (§ 6.15) (cp. Vining et al., 1986, p. 681).

Although no linear relationships were indicated between OAR and age, OAR and education, and age and education, there is no reason to believe that curvilinear or otherwise systematic nonlinear relationships exist between these variables.[23]

Partial Correlations r_p. The partial correlation coefficient (r_p) is related to the Pearson correlation coefficient (r) and as a result can be treated similarly (Green et al., 1997, p. 294; Erickson & Nosanchuk, 1977, pp. 340, 343). Partial correlations were calculated in order to identify possible confounding effects of the identified intervening variables (i.e., age, education, and Police Intake Exam score), essential for interpreting any correlation that may be found between OAR and sex. Specifically, once the bivariate correlation between the criterion variable (i.e., OAR) and the predictor variable (i.e., sex) was known, which is the correlation of interest, partial correlations ("pairwise" and "listwise") were calculated to remove the effects of the control variables, separately (first-order) and collectively (second-order and third-order) (Miller &Whitehead, 1996, p. 327; SPSS® Base 10.0, 1999, pp. 178-179). As a result, it was possible to estimate the correlation between OAR and sex as if all candidates were of the same age, education, and mental ability.[24] In addition, partial correlations were (1) tested for statistical significance (one-tailed), where p values were reported (along with Bonferroni adjustments); (2) squared (r_p^2) to interpret the strength of the relationships; and (3) corrected for attenuation.

---

[23] For those correlations where no linear relationship is indicated, a visual inspection of scatterplots created for each pair of variables (see Appendix 3) does not indicate any obvious systematic nonlinear relationships (rf. Erickson & Nosanchuk, 1977, p. 241; Green et al., 1997, p. 282; Miller and Whitehead, 1996, pp. 319-321; and SPSS® Base 10.0, p. 178). Note that scatterplots were not created for pairs that include the categorical variable of sex.

[24] It is assumed that mental ability or intelligence is measured by the Police Intake Exam score.

Assumptions for bivariate and partial correlations are the same, except that for partial correlations there is the assumption that the variables are multivariately normally distributed (Green et al., 1997, p. 293). That is, each variable is independently normally distributed, and each variable is normally distributed for all levels of the other variable combinations, which again results in a linear relationship between the variables of interest. The assumption of linearity, however, does not have to be met perfectly, since linear regression is robust (partial correlation is based on regressions, where residuals are correlated), especially if the sample size is large (Erickson & Nosanchuk, 1977, pp. 241, 340, 343).

3. Binomial effect-size display (BESD). The correlation coefficient is often considered an index of effect size (Green et al., 1997, p. 283; Rosnow & Rosenthal, 1996, p. 255); and particular attention is paid to statistical significance or *p* values, which is in part a function of the number of cases in the sample (Erickson & Nosanchuk, 1977, p. 241; Miller & Whitehead, 1996, p. 326; Rosnow & Rosenthal, p. 262). However, despite coefficient size and statistical significance, there is the question of practical significance or importance. For example, deceptively small correlation coefficients may have practical significance, and statistically significant *p* values may have no practical significance.[25] As a result, the BESD procedure described by Rosnow & Rosenthal (pp. 257-263, citing Rosenthal & Rubin, 1982) was used to translate the correlation coefficient between OAR and sex into dichotomous outcomes, which were then presented in a tabular display to evaluate the practical importance of *r* (see also Rosenthal &

[25] As stated by Erickson and Nosanchuk (1977), "If *N* is large enough then almost any weed of a relationship will turn up as significant" (p. 241). For a discussion within the context of discrimination, see Vining et al. (1986, p. 681), and Paetzold and Willborn (1999, §§ 4.13, 4.15).

Rosnow, 1991, pp. 280-285). As sex is naturally dichotomous, it was only necessary to transform OAR into a dichotomous variable, which was split at the median (38); i.e., below 38 was considered a fail, while 38 and above was considered a pass, which is consistent with how police managers interpret the results.

CHAPTER VI: RESULTS AND DISCUSSION

As part of the overall goal of providing a model to assess systemic discrimination in an assessment center, this study posed two general research problems. First, addressing the issue of measurement error, whether the Police Academy assessment center meets conventionally accepted standards for interrater reliability. And second, within the framework for a quantitative analysis of discrimination, whether the Police Academy assessment center discriminates on the prohibited ground of sex. The purpose of this chapter is to report and discuss the results of the study.

Interrater Reliability

Specifically, the research question was whether the relative consistency (i.e., correlation) amongst Police Academy assessment center raters meets the conventionally accepted standard of .60. To answer this question, this section will report, separately (by year) and collectively (1978-1999), the interrater reliability coefficients for the sample, which consists of 55 police recruit Assessment Centers (ACs). Because, reliability was estimated by three different techniques (the correlational approach, Cronbach's alpha, and intraclass), it was possible to make comparisons to determine if the correlational approach produced results similar to that of the more complex alpha and intraclass techniques. Finally, based on the collective sample, the results of an item (dimension) analysis and internal consistency (scale) analysis will be reported.

An Assessment of Interrater Reliability

Table 6-1 below shows average reliability coefficients over a period of 22 years at the Police Academy assessment center.

Table 6-1

Mean[a] interrater reliability[b] (fixed raters), summarized by year

		Type			
Year	N (ACs)	Corr R^c	Alpha α^d	$ICC_{3,1}$ r^e	$ICC_{3,k}$ R^f
1978	1 (1)	.50	.69	.31	.69
1979	.	.	.	.	.
1980	3 (4)	.80	.76	.41	.76
1981	3 (7)	.62	.68	.33	.68
1982[g]	1 (8)	.13	-.41	-.08	-.41
1983	1 (9)	.47	.46	.18	.46
1984	2 (11)	.52	.52	.25	.52
1985	1 (12)	.05	.06	.02	.06
1986	1 (13)	.41	.08	.02	.08
1987	2 (15)	.59	.61	.35	.61
1988	4 (19)	.80	.74	.45	.74
1989	3 (22)	.35	.35	.18	.35
1990	5 (27)	.76	.74	.45	.74
1991[g]	4 (31)	.49	.06	.40	.06
1992	4 (35)	.45	.54	.26	.54
1993	1 (36)	.39	.39	.14	.39
1994[g]	2 (38)	.18	.44	.24	.44
1995	3 (41)	.66	.62	.32	.62
1996	3 (44)	.91	.87	.73	.87
1997[g]	4 (48)	.47	.33	.22	.33
1998	4 (52)	.84	.79	.63	.79
1999	3 (55)	.74	.72	.46	.72
Grand Mn[h]	n/a	.64	.59	.36	.59

Notes:

[a] Means computed using Fisher's *Z* transformation.

[b] For overall assessment ratings (OARs).

[c] Corr *R*: Correlational approach, corrected (equivalent to standardized item alpha).

[d] Alpha α: Cronbach's (true) alpha.

[e] $ICC_{3,1}$ *r*: Intraclass correlation coefficient, configuration #3, single rater.

[f] $ICC_{3,k}$ *R*: Intraclass correlation coefficient, configuration #3, corrected.

[g] Some ACs had negative reliability coefficients, which often produces irregular results.

[h] Weighted.

. indicates no data available.

As indicated, the average overall interrater reliability for a group of raters (i.e., corrected reliability), depending on the technique used, was estimated between .64 and .59, while the average reliability for a single rater was estimated at .36. Notably, negative reliability coefficients for individual ACs were not adjusted to 0.00, and as a result the "true" overall average reliability may be slightly underestimated. Theoretically, reliability is bounded by zero and one (as discussed in Chapter 3), where .00 indicates no reliability and 1.00 indicates perfect reliability (Carmines & Zeller, 1979, pp. 31, 45; Tinsley & Weiss, 1975, p. 363; Traub, 1994, p. 38).

Though theoretically meaningless, negative reliability coefficients are mathematically possible when using techniques such as Pearson's product-moment correlation, Cronbach's alpha, or intraclass. For example, configuration #3 (corrected for a set of raters) of the intraclass technique uses analysis of variance computations, which yield the between candidates mean square and the residual mean square. Here, if the residual mean square exceeds the between candidates mean square, once applied to the appropriate formula ($ICC_{3,k}$)[1] the result is a negative reliability coefficient (see, for example, Table 6-1, 1982).[2] Moreover, the results can be quite irregular, as found in 1991, where the reliability for a single rater was higher than for a set of raters.

Although the reliability coefficients were calculated for fixed sets of raters, the results may be assumed to represent interrater reliability in general at the Police Academy assessment center as the sample of 55 ACs was randomly selected. Notably, when a random effects model was used for the collective sample, which allows the results to be

[1] Refer to Chapter 3, where the theory and formulae are described in detail.
[2] According to Tinsley and Weiss (1975), negative reliability coefficients indicate possible interaction between raters and candidates (p. 363).

generalized to a population of raters, similar reliability coefficients (.56 for a set of raters and between .24 and .30 for a single rater) were found, as shown in Table 6-2 below.

Table 6-2

Interrater reliability[a] (random raters)

			Type	
Raters	Year	N (ACs)	$ICC_{1,1}$ r[b]	$ICC_{1,k}$ R[c]
3	1978-99	55	.30	.56
4	1978-99	55	.24	.56

Notes:
[a] For overall assessment ratings (OARs).
[b] $ICC_{1,1}$ r: Intraclass correlation coefficient, configuration #1, single rater.
[c] $ICC_{1,k}$ R: Intraclass correlation coefficient, configuration #1, corrected.

Comparing the results reported in Tables 6-1 and 6-2 with the conventionally accepted standard of .60 (see Chapter 3, Table 3-8), the average overall reliability at the Police Academy assessment center appears to be adequate, regardless of the technique, although the correlational approach results in a slightly higher estimate. Nevertheless, an inspection of Table 6-1 indicates a few unacceptably low corrected reliabilities (i.e., $R <$ .30), along with some substantially inconsistent estimates. Specifically, average corrected reliability estimates (R and α respectively) below .30 were found in 1982 (.13, -.41), 1985 (.05, .06), 1986 (.41, .08), 1991 (.49, .06), and 1994 (.18, .44).

There are, however, two possible explanations for these anomalies: negative reliability coefficients and range restriction. In 1991 and 1994 there were individual ACs that had negative reliability coefficients, which contributed to the irregular results noted above. When these negative coefficients are adjusted to zero, the average corrected reliability estimates (R and α respectively) are .60 and .62 for 1991, and .46 and .46 for 1994. For the remaining years (1982, 1985, and 1986), range restriction, a limitation of

parametric indices that rely on variability of scores (such as the indices used in this study), may have contributed to low reliabilities.[3]

Evidence of range restriction is found in the fact that the Police Academy's rating scale has thirteen possible ratings (see Table 5-4), but only seven ratings (converted ratings between 22 and 42) accounted for 97% of all ratings. As a result, restricted standard deviations for candidate OARs were found in the collective sample (sd = 4.56) and in 1982 (sd = 5.91), 1985 (sd = 0.00), and 1986 (sd = 1.55),[4] which can produce artificially low reliabilities (Cronbach, 1984, pp. 172-173; Whitehurst, 1985, p. 568; Guilford & Fruchter, 1978, p. 431). For example, in 1985, for the class selected, the raters all agreed on an OAR of 38 for all six candidates (i.e., there was perfect agreement); however, because there was almost no variability in the original rater assessments the reliability was close to zero.[5] For 1986 (where alpha and intraclass resulted in a corrected reliability of .08), after correcting for range restriction by using the formula presented in Guilford and Fruchter (1978),[6] using the collective standard deviation of 4.56 as the more variable standard, reliability was estimated at .49. The results found in 1982 will be discussed later in this chapter.

Finding range restriction at the Police Academy assessment center was consistent with the results of a meta-analysis conducted by Gaugler, Rosenthal, Thornton and Bentson (1987), who found that assessment centers show moderate to severe range

---

[3] See Chapter 3, where range restriction was discussed in detail in a separate section.
[4] For the purposes of comparison, it is useful to note that if all 13 ratings were used and variability was maximized across categories, the standard deviation would be 12.50.
[5] The interrater reliability was not exactly zero because, as discussed in Chapter 5, the OAR reflects a consensus score, which is not necessarily identical to the original assessments given by the individual raters.
[6] See Chapter 3 for details.

restriction (p. 494). A plausible explanation is that of group homogeneity, where candidates are very similar on the dimensions of interest. Because of the expense, only the most promising candidates are nominated to attend assessment centers, and so homogeneity on the dimensions of interest is likely. However, an alternate explanation is that of systematic rating error, which includes errors of central tendency, leniency, and halo. Although it is impossible to eliminate rating error as a cause, it seems reasonable to conclude that group homogeneity was contributing in part to range restriction at the Police Academy assessment center.

The Correlational Approach Compared to Alpha and Intraclass

Arranged in a stem and leaf display, Table 6-3 classifies the average reliabilities, which are grouped by technique, for each year according to conventionally accepted standards (rf. Table 3-8), and so is useful for comparing the correlational approach with Cronbach's alpha (and intraclass).

Table 6-3

Adequacy of reliabilities, by year

Reliability	Adequacy	Type	Year (1978-1999)
< .30	less than	Corr R^a	19:82 85 94
		Cron α^b	19:82 85 86 91
≥ .30 < .45	marginal	Corr *R*	19:89 93
		Cron α	19:89 93 94 97
≥ .45 < .60	low	Corr *R*	19:78 83 84 86 87 91 92 97
		Cron α	19:83 84 92
≥ .60 < .75	acceptable	Corr *R*	19:81 95 99
		Cron α	19:78 81 87 88 90 95 99
≥ .75 < .90	high	Corr *R*	19:80 88 90 98
		Cron α	19:80 96 98
≥ .90	very high	Corr *R*	19:96
		Cron α	19:

Notes:
[a] Correlational approach, corrected.
[b] Cronbach's (true) alpha, which is equivalent to intraclass *R* (configuration #3, corrected ($ICC_{3,k}$)).

An examination of this table shows that even though the correlational approach yielded a slightly higher overall average reliability coefficient (.64), when individual years (see Table 6-1) were classified according to conventional standards, the results were similar to Cronbach's alpha. For the correlational approach, 13 out of 22 years were classified as either low, marginal, or inadequate, and for Cronbach's alpha (and the intraclass technique), which yielded a lower overall average reliability (.59), 11 years were classified as low, marginal, or inadequate.

Of particular importance is that when reliabilities are positive, differences between the correlation approach and Cronbach's alpha and intraclass are small and generally not significant at a practical level. As shown in Table 6-1, large differences between techniques were found in 1978, 1982, 1986, 1991, 1994 and 1997; however, negative reliability coefficients produced irregular averages for 1982, 1991, 1994 and 1997. By adjusting these negative coefficients to zero, the average reliabilities for the correlational approach and Cronbach's alpha, respectively, were .13 and .00 for 1982, .60 and .62 for 1991, .46 and .46 for 1994, and .50 and .44 for 1997. However, without adjusting for negative reliability coefficients or for range restriction, the overall (1978-1999) difference between the correlational approach and Cronbach's alpha was only .05, which is consistent with that reported in the literature (Cronbach, 1984, pp. 169-170; Fleenor, Fleenor & Grossnickle, 1996, pp. 377-378; Rosenthal & Rosnow, 1991, pp. 55-56).

Item Analysis and Internal Consistency Analysis

Reliability analysis, in addition to providing summary measures of interrater reliability, can be used to isolate areas of the assessment center which require

improvements if reliability in general were to be enhanced. Table 6-4[7] shows the results of the item (i.e., dimension) analysis and internal (i.e., scale) consistency analysis, along with Cronbach's alpha (rf. Hinrichs & Haanpera, 1976, pp. 38-39),[8] all of which are discussed in detail below.

Table 6-4

Item analysis (rank order) and internal consistency analysis

Dimensions	IA[a]	ICA[b]	Dimensions	IA[a]	ICA[b]
Personal Impact	.79	.57	Fact Finding	.60	.44
Maturity	.77	.50	Able to Learn	.59	.43
Problem Confrnt.	.76	.53	Tolerance[c]	.56	.42
Stress Tolerance	.75	.53	Interpersonal Sens.	.53	.35
Initiative	.72	.49	Integrity	.49	.27
Oral Commun.	.72	.54	Written Commun.	.28	.17
Intelligence	.72	.52	Adhere to Auth.[c]	.28	.17
Decisiveness	.66	.46	Flexibility	.14	.09
Cronbach's true alpha (α)[d]				n/a	.91[e]
Cronbach's standardized item alpha (α)[d]				n/a	.90[e]

Notes:
[a] Item Analysis: correlation (r) between dimension and OAR.
[b] Internal Consistency Analysis: average correlation ($\bar{r}$) between dimension and other dimensions, excluding dimensions 15 and 16. All averages computed using Fisher's Z transformation.
[c] Average correlation between dimension and all other dimensions.
[d] N = 289. Does not include dimensions 15 and 16, which were discontinued in 1994 (discussed later in this section).
[e] Alpha necessarily exceeds the average correlations for each dimension (which are of concern individually) because alpha is corrected upwards by a factor of 14 according to the Spearman-Brown formula (discussed in Chapter 3).

Item Analysis

The purpose of the item analysis was to determine whether individual items (i.e., dimensions) were making a meaningful contribution to overall candidate assessments.

[7] Table 6-4 is a summary of (1) all bivariate correlations between the final ratings for each dimension and OAR and (2) the average bivariate correlations between dimensions. For a table showing all bivariate correlations, see Appendix 4.
[8] An item analysis can be very sophisticated; however, for the purposes of this study, the methodology suggested by Hinrichs and Haanpera (1976) is sufficient.

Although the cut point is somewhat arbitrary, Hinrichs and Haanpera (1976) suggest that for assessment centers correlations (r) of less than .20 indicate a "lack of contribution" to overall assessments (p. 38). With .20 as a general standard, the dimension of flexibility ($r = .14$) was not making a contribution to overall assessments. Flexibility is admittedly an important dimension, as the use of discretion is integral to policing in a democracy (e.g., see Hooke, 1996; Kleinig, 1996; and Walker, S., 1993); but if this dimension is to be useful it needs to be reviewed and redefined and raters need to be retrained, after which it needs to be periodically reassessed and modified as necessary.

The dimension of written communications, although above .20, was also not contributing well ($r = .28$) to overall candidate assessments. As with flexibility, written communication is important to successful police work, but it is defined by more technical skills and so seems out of place with the other dimensions. Notably, written communications was the only dimension where candidates completed an assignment alone and unobserved, while the remaining dimensions focussed on candidate interaction with role players in the various simulations. Therefore, it appears that written communications was measuring an attribute not measured by the other dimensions, and so should be tested separately, much as the tests for driving and physical abilities.

For reasons unknown, the dimension of adherence to authority was discontinued by the Police Academy in 1994 and so it requires no discussion except to note that it was contributing little to overall candidate assessments.

Internal Consistency Analysis

Analyzing internal consistency is also useful for making improvements, as dimensions (analogous to items in a test) that do not correlate well with other dimensions

decrease overall reliability on a scale that is intended to be unidimensional (Hinrichs & Haanpera, 1976, p. 38).[9] Of particular importance to assessment centers, however, is that dimensions may be independent sub-scales (measuring distinct attributes) of a larger scale that is intended to measure an overall construct (e.g., basic competencies of a successful police officer). Consequently, because a dimension has a low average correlation with other dimensions does not necessarily mean that it is deficient. Rather, if the dimension is highly correlated to the overall assessment, it may mean that the dimension has discriminant validity (see Chapter 2).

As in the item analysis discussed above, although deciding on a cut point is somewhat arbitrary, Hinrichs and Haanpera (1976) suggest that an average dimension correlation ($\bar{r}$) less than .30 is low (p. 39). And so, with .30 as a general standard, the dimension of flexibility ($\bar{r}$ = .09) was not correlating well with the other dimensions. As previously discussed, flexibility is an important dimension, but its lack of correlation with the other dimensions and its lack of contribution to overall assessments clearly indicate its needs for revision and reassessment.

The average correlation for integrity ($\bar{r}$ = .27) fell below the standard of .30, but its correlation (r = .49) with overall assessments was adequate. Any interpretation of these results is confounded by historically inconsistent scoring procedures for this particular dimension (Sergeant MacDonald, personal communication, November, 1999). For example, at times candidates received either a converted score of 35 (a pass) (see Table 5-4) or no score (a fail), or converted scores between 32 and 38 or no score.

[9] Refer to the discussion in Chapter 3 on the history of reliability theory, where Spearman and Brown in 1910 proved that reliability increases as test items are increased, assuming that all items are positively correlated.

However, despite these scoring problems, it is possible that the low average correlation with the other dimensions is the result of raters clearly discriminating between this dimension and the others (i.e., it has high discriminant validity). This explanation is supported by the fact that its contribution to the overall assessment was adequate and that both the Police Academy and nominating police departments generally view integrity as a unique or special dimension. Candidates who failed the integrity dimension, regardless of their standing in the other dimensions, were usually rejected by their nominating police departments. As a result, that integrity did not correlate well, on average, with the other dimensions was not unexpected.

Similar to integrity, the average correlation for the dimension of interpersonal sensitivity was moderately low ($\bar{r} = .35$), but its correlation ($r = .53$) with overall assessments was adequate. Again, it is possible that this moderately low average correlation (compared to the standard of .30) is the result of raters clearly discriminating between interpersonal sensitivity and the other dimensions (i.e., it has high discriminant validity). This explanation is supported by the fact that its contribution to the overall assessment was adequate and that assessment centers are necessarily interactive, where interpersonal sensitivity may be easier to distinguish from the other dimensions.

The average correlation of written communications was low ($\bar{r} = .17$) as was its correlation with overall assessments, which corroborates the suggestion that it should be treated separately. Notably, the dimension of interpersonal tolerance was discontinued in 1994 because it was believed to be replicating interpersonal sensitivity (Sergeant MacDonald, personal communication, November, 1999), which appears to be confirmed by the correlation between these two dimensions ($r = .73$) (see Appendix 4). Finally, the

average correlation for the dimension of adherence to authority, which for unknown reasons was also discontinued in 1994, was low ($\bar{r} = .17$). Considering that this dimension also contributed little to overall assessments ($r = .28$), and that by definition it appeared to be in conflict with the definition of flexibility (see Table 5-2), it appears that this dimension had little to contribute on any level.

Cronbach's Alpha

Cronbach's alpha (α) is a good summary measure of internal consistency (see Chapter 3) for a scale or sub-scale that is intended to be unidimensional (i.e., one that measures a single construct or attribute). Specifically, a low alpha may indicate that the scale is unreliable; but, apart from reliability, it may also indicate that the scale is composed of sub-scales that are measuring disparate attributes (Cronbach, 1951, pp. 320, 331; Cronbach, 1970, pp. 300, 331; Guilford & Fruchter, 1973, p. 407; Kaplan & Succuzzo, 1982, p. 103).

The coefficients for Cronbach's alpha, as shown in Table 6-4, were .91 and .90 for true alpha and standardized item alpha, respectively, and are very high by conventional standards (rf. Rosenthal and Rosnow, 1991, pp. 50-51).[10] It appears, then, that the Police Academy recruit assessment center scale is unidimensional, which seems contradictory given that it has 14 dimensions that were designed to measure different attributes. However, such a finding is possible if there is considerable overlap between the dimensions and if the dimensions lack discriminant validity. The significance of this

[10] Standardized item alpha is based on the average inter-item correlation, corrected by the Spearman-Brown formula (i.e., it is equivalent to correlational *R*), and is often similar to true alpha as shown in Table 6-1. Because it is corrected, it is necessarily greater than the average of all inter-item correlations.

finding is that it questions the traditional dimension-based theory,[11] where it is assumed that raters can adequately differentiate between dimensions, each of which are designed to measure separate attributes of some overall construct of interest (e.g., competency as a police officer).

Summary and Conclusions

Compared with the conventionally accepted standard ($R = .60$), the average overall reliability at the Police Academy assessment center appears to be adequate, regardless of the technique used. Although unacceptably low reliabilities ($R < .30$) were found in 1982, 1985, 1986, 1991, and 1994, along with some substantially inconsistent estimates, both anomalies were generally explained by negative reliability coefficients and range restriction, except for 1982, where low reliability cannot be explained away.

Overall, when compared to Cronbach's alpha and intraclass, the correlational approach slightly overestimated corrected reliability (.64 compared to .59), but not to an extent that would likely mislead administrators of operational assessment centers. Moreover, when used to estimate reliability according to the classifications suggested in Table 6-3, both the correlational approach and Cronbach's alpha yielded similar results. Operationally, then, the correlational approach appears to be an acceptable alternative to the more complicated Cronbach's alpha or intraclass technique.

To ensure that an assessment center is operating efficiently and that benefits are maximized, a reliability analysis should include an item analysis and an internal consistency analysis (Hinrichs & Haanpera, 1976, pp. 35, 38). These analyses indicated

[11] See Chapter 2 for a comprehensive discussion of discriminant validity and the debate between dimension-based theory and exercise-based theory.

that the dimension of flexibility needed a comprehensive review and that the dimension of written communication was measuring a skill set completely unrelated to the other dimensions. Overall internal consistency, as estimated by Cronbach's alpha, was high (.91 and .90 for true and standardized alpha, respectively), which suggests that, except for integrity and interpersonal sensitivity, the dimensions overlap and may lack discriminant validity.

Systemic Sex Discrimination

This section addresses the issue of systemic sex discrimination at the Police Academy assessment center; and more specifically, answers the question of whether or not there is a correlation between sex and OAR (i.e., whether OAR is a function of candidate sex). This question, however, must be interpreted within a legal framework; i.e., whether the evidence, if admissible, is sufficient to convince the fact-finder that discrimination has occurred and that it is of some practical significance. Therefore, following the model proposed for this study, this section represents the second (assessment) phase, the purpose of which is to report and discuss the results of the parity and probity analyses.

Parity (Comparative Evidence)

Table 6-5 divides OARs, grouped by sex, into three time periods: 1978 to 1988, 1989 to 1999, and 1978 to 1999. Regardless of the time period, the results show that female candidates consistently received higher OARs than male candidates (overall, 37.95 and 36.55 respectively). Table 6-6 displays OARs by rating categories, where an OAR of less than 38 (the median score) was generally considered a failing score (this is

discussed in more detail in the next section). The results show that males consistently received a higher proportion of scores in each OAR rating category under the median, while females consistently received a higher proportion of scores in each rating category above the median (with an overall difference of 13.2%).

Table 6-5

OARs: Comparing males with females (1978-1999)

Years	Sex	N	Mn	Sd	Min	Max
1st 11 years: 1978-1988	Male	988	36.38	4.61	15	48
	Female	90	38.42	4.59	25	48
	Total	1078	36.55	4.64	15	48
2nd 11 years: 1989-1999	Male	1480	36.66	4.10	15	48
	Female	383	37.84	3.58	25	45
	Total	1863	36.90	4.02	15	48
Total: 1978-1999	Male	2468	36.55	4.31	15	48
	Female	473	37.95	3.79	25	48
	Total	2941	36.78	4.26	15	48

Table 6-6

Ratings distribution (1978-1999), grouped by sex

OARs	Male		Female		Total		% Differ.
Fail (≤ 35)	N	%	N	%	N	%	M – F =
15 (very poor)	5	.2	0	.0	5	.2	+ .2
18 (poor)	0	.0	0	.0	0	.0	.0
22 (very little ability)	3	.1	0	.0	3	.1	+ .1
25 (well below average*)	23	.9	2	.4	25	.9	+ .5
28 (below average)	188	7.6	10	2.1	198	6.7	+ 5.5
32 (slightly below average)	292	11.8	48	10.1	340	11.6	+ 1.7
35 (average/competent)	535	21.7	78	16.5	613	20.8	+ 5.2
Total	1046	42.3	138	29.1	1184	40.4	+ 13.2
Pass (≥ 38)	N	%	N	%	N	%	M – F =
38 (slightly above average)	930	37.7	197	41.6	1127	38.3	- 3.9
42 (above average)	443	17.9	115	24.3	558	19.0	- 6.4
45 (well above average)	47	1.9	22	4.7	69	2.3	- 2.8
48 (a great deal of ability)	2	.1	1	.2	3	.1	- .1
52 (superior ability)	0	.0	0	.0	0	.0	.0
55 (excellent)	0	.0	0	.0	0	.0	.0
Total	1422	57.6	335	70.8	1757	59.7	- 13.2
Grand Total	2468	100.0	473	100.0	2941	100.0	.0

Note: Minor discrepancies are due to rounding at the first decimal.

* The Police Academy uses the term "average," but Academy documentation clearly indicates that the rating scale is criterion-referenced (i.e., 25 should be defined as "well below competent").

Therefore, because the differences in OARs appeared to be systematic and sex based (i.e., the pattern of scores suggests that the Police Academy assessment center may have systematically discriminated against males), a probity analysis was conducted to determine if the differences were legally significant.

Probity (Legal/Practical Significance)

The differences in scores indicated in the parity analysis were analyzed by means of the four-fifths rule, statistical tests, and the binomial effect-size display.

The Four-fifths Rule

The four-fifths rule is conceptually similar to traditional statistical testing procedures, as both provide the basis for making inferences about discrimination. According to the four-fifths rule, if the success rate (i.e., selection or pass rate) for any sub-group within a pool of candidates falls below .80 of the dominant sub-group, a prima facie case for discrimination has been made. Table 6-7A shows the adverse effect rate when the cut score for passing was set at an OAR of 35, which the assessment center's rating scale (see Table 3-1) defines as "average" or "competent."

Table 6-7A

Pass = 35: Adverse effect rate (1978-1999)

	Male		Female		Adverse Effect	
OAR	N	%	N	%	F/M	%
Fail (≤ 32)	511	20.6	60	12.6	n/a	n/a
Pass (≥ 35)	1957	79.3	413	87.3	M*	90.8
Total	2468	100.0	473	100.0	n/a	n/a

Notes:
Minor discrepancies are due to rounding at the first decimal.
n/a indicates not applicable.
* indicates sub-group that was not dominant.

Although the results show that the overall male pass rate was only 79.3% while the female pass rate was 87.3%, the adverse effect rate was 90.8%, which well exceeds the four-fifth's threshold of 80%. However, according to Sergeant MacDonald, the Police Academy assessment center administrator, the cut score of 35 is inconsistent with how the OAR has been applied operationally by the sponsoring police departments (personal communication, November 1999). Police personnel managers generally consider that a candidate has failed the assessment center if his or her OAR is less than 38 (or 3+),[12] which the rating scale defines as "slightly above average." Consequently, it was more appropriate to use 38 as the cut score (which was also the median score), and the results are shown in Table 6-7B.

Table 6-7B

Pass = 38: Adverse effect rate (1978-1999)

	Male		Female		Adverse Effect	
OAR	N	%	N	%	F/M	%
Fail (≤ 35)	1046	42.3	138	29.1	n/a	n/a
Pass (≥ 38)	1422	57.6	335	70.8	M*	81.4
Total	2468	100.0	473	100.0	n/a	n/a

Notes:
Minor discrepancies are due to rounding at the first decimal.
n/a indicates not applicable.
* indicates sub-group that was not dominant.

Here, the difference between pass rates was greater, where the overall male pass rate was only 57.6% compared to the overall female pass rate of 70.8%, but the adverse effect rate of 81.4% was still greater than the threshold of 80%.[13] Therefore, according to

[12] In most cases, candidates who received an OAR of less than 38 were (and are) eliminated from the recruitment process; therefore, to use a cut score of 35 would artificially minimize the "adverse effect," which defines the substantive approach used by the Supreme Courts in both Canada and the U.S. Consequently, for the purposes of this study, an OAR of less than 38 was considered a failing score.
[13] If the cut score were 42, then the adverse effect rate against males would be approximately 68%. Proportionately, as indicated in Table 6-8, more females received higher scores than did males.

the four-fifths rule per se, there is insufficient evidence to support an inference that the Police Academy assessment center has discriminated on the basis of sex (i.e., favoring females over males or males over females). That said, it is important to recognize that the four-fifths rule has limitations, such as its inherent arbitrariness. However, the problem of arbitrariness, for which there is no easy solution, also applies to the traditional *p* value of .05, which is the significance threshold for most statistical testing.[14]

Another limitation is that the four-fifths rule is insensitive to sample size. For example, if less than 200 candidates are selected an adverse inference is more likely under the four-fifths rule,[15] if 200 are selected an adverse inference is equally likely under either the four-fifths rule or traditional statistical tests of significance, and if more than 200 are selected an adverse inference is more likely under traditional tests of significance (Vining, McPhillips & Boardman, 1986, pp. 689-691). This limitation is recognized by the Equal Employment Opportunity Commission (EEOC), which notes that small differences in selection rates that do not violate the four-fifths rule may constitute "adverse impact"[16] when the differences are significant in both statistical and practical terms. On the other hand, the EEOC notes that large differences in selection rates that do violate the four-fifths rule may not constitute "adverse impact" when they are based on small numbers and are not statistically significant (cited in Paetzold & Willborn, 1999, §§ 5.06, 5.07).

---

[14] Paetzold and Willborn (1999) argue that the value of the four-fifths threshold of .80 is limited because it is only accepted in discrimination litigation and related studies, while the traditional *p* value of .05 is not only accepted in discrimination litigation and related studies but it is also accepted in the social sciences, where it has gained almost universal recognition (§ 5.06).

[15] Along with making a Type I error; i.e., rejecting the null hypothesis of no discrimination in favor of the alternative when the null hypothesis is in fact true.

[16] "Adverse impact" is a U.S. term that has the same general meaning as the Canadian terms of "adverse effect" or "discriminatory effect."

It follows, then, because the main sample for this study was a large pool of candidates (N = 2956), where the sub-groups were substantially greater than 200, that the difference in pass rates may yet be significant both in statistical and practical terms, where an inference of adverse or discriminatory effect may be made. Accordingly, this study now turns to the results of traditional statistical tests, which were used to supplement the findings of the four-fifths rule (rf. Paetzold & Willborn, 1999, § 5.07; see also the EEOC's recommendations noted above).

Statistical Tests

A standard *t*-test (independent means) comparing the combined total mean OAR (36.55) of males with the combined total mean OAR (37.95) of females (rf. Table 6-5) indicated that the difference between male and female OARs was statistically significance (p (t_{2939} = -6.588) < .000 (one-tailed, equal variances assumed)). To infer a "discriminatory effect" against males, however, may be premature because of the problem of identifying groups that are appropriate for a comparative analysis and how they should be analyzed, and because of the problem of determining practical significance.

As discussed in Chapter 5, the appropriate population for analysis is the existing pool of candidates; however, these candidates must also be "qualified," which is itself a particularly contentious issue in the discrimination debate (Vining et al., 1986, p. 682). For example, although an employer may consider a particular level of education, such as a university degree, to be a necessary job prerequisite, it may not meet the rigorous standards of a bona fide occupational requirement (BFOR). Nevertheless, the identification of all screening criteria (along with general demographic data) is essential

to isolate those factors that may differentially contribute to the selection process (Paetzold & Willborn, 1999, § 4.16; Vizkelety, 1987, p. 185).

With this in mind, it is noteworthy that the analysis of the candidate pool in Chapter 5 indicated that the female group was, on average, older (rf. Table 5-8), better educated (rf. Table 5-9), and scored higher on the Police Intake Exam (rf. Table 5-10).[17] As age, education and Police Intake Exam score may be good predictors of assessment center performance, and because they appear to be related to candidate sex, as a prerequisite to a discrimination inference it was necessary to explore the possibility that these variables, rather than sex, were contributing to the differences in OARs. Because univariate statistical tests are generally inadequate to accommodate multiple variables (although the candidate pool can be stratified or groups appropriately matched),[18] two relatively simple multivariate statistical tests (bivariate[19] and partial correlations) were selected to address this problem.

Pearson's product-moment correlation *r*. Similar to the parity analysis, the combined sample was divided into three categories: 1978 to 1988, 1989 to 1999, and 1978 to 1999. As indicated in correlation matrix displayed in Table 6-8, the correlations between OAR and sex were almost identical (.122, .118, and .121 respectively). A standard *t*-test (independent means) comparing the total mean OAR (36.55) of the first 11

---

[17] As discussed in Chapter 5, it is assumed that the Police Intake Exam is measuring, to an unknown degree, mental ability, which is here defined as intelligence, cognitive ability or academic skills or some combination thereof.

[18] In the U.S., the binomial probability distribution test (or sometimes the non-parametric chi-square), which treats data categorically, is often used (Paetzold & Willborn, 1999, § 4.11, citing Hazelwood School Dist. v. United States, 1977, as an example). Similar to a univariate *t*-test, this technique requires the stratification of aggregate data to isolate variables such as education, which is not as precise as a multivariate technique. More will be said about this issue later in the chapter.

[19] Technically, a bivariate correlation is not multivariate, which is defined as involving more than two variables (Kachigan, 1991, p. 142). However, bivariate correlations are useful for exploring the relationships amongst several variables, similar to correlational analyses in general.

years with the total mean OAR (36.90) of the second 11 years (see Table 6-5) indicated that the difference between OARs was not statistically significant at .01, but was significant at .05 (p (t_{2944} = -2.115) = .035 (two-tailed, equal variances assumed)). Considering the large sample sizes,[20] the results suggested that a further analysis of scores stratified by time was not necessary.

Table 6-8

OAR correlation matrix (1978-1999): Pairwise[a]

Years	Details	N	Sex	Age	Educ.	Exam
1st 11 years: 1978-1988	OAR	1.000	.122	.184	.025	.221
	r^2	.	.015	.034	.001	.049
	p (2-tail)[b]	.	.000[c]	.000	.555	.077
	N (valid)	1078	1078	596	546	65
2nd 11 years: 1989-1999	OAR	1.000	.118	.144	.130	.194
	r^2	.	.014	.021	.017	.038
	p (2-tail)[b]	.	.000[c]	.001	.002	.016
	N (valid)	1868	1863	570	558	154
Total: 1978-1999	OAR	1.000	.121	.176	.102	.222
	r^2	.	.015	.031	.011	.049
	p (2-tail)[b]	.	.000[c]	.000	.001	.001
	N (valid)	2946	2941	1166	1104	219

Notes:
[a] In a "pairwise" selection, sample size varies from pair to pair.
[b] In U.S. discrimination jurisprudence, statistical significance is generally accepted when $p < .05$ (similar to the social sciences). With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0125 (.05/4).
[c] p (1-tail).

Consistent with the results of the t-test comparing males with females, the correlation between OAR and sex (.121) was statistically significant ($p < .000$). As previously hypothesized, though, OAR was also correlated to age, education, and Police Intake Exam score (.176, .102, and .222 respectively), and these correlations were also statistically significant (see Table 6-8). Notably, correlation sizes varied from slight to

[20] Statistical significance is in part a function of sample size; therefore, because the sample sizes were quite large, an alpha level of .05 would probably be too low to assume statistical significance.

low, with statistical significance in part a function of large sample size. This fact was highlighted by the coefficients of determination (r^2), which indicated that the proportion of shared variances ranged from only 1.5% to 4.9%.

Table 6-9, building on correlation matrix of Table 6-8, gives the bivariate correlation coefficients between all variables of interest.

Table 6-9

General correlation matrix (1978-1999): Pairwise[a]

	1 OAR	2 Sex	3 Age	4 Educ.	5 Exam
1 OAR	1.000	.121	.176	.102	.222
r^2	.	.015	.031	.011	.049
p (2-tail)[b]	.	.000[c]	.000	.001	.001
N	2946	2941	1166	1104	219
2 Sex		1.000	.084	.175	.161
r^2		.	.007	.031	.026
p (2-tail)		.	.004	.000	.017
N		2951	1166	1104	219
3 Age			1.000	.069	.334
r^2			.	.005	.112
p (2-tail)			.	.021	.000
N			1166	1101	218
4 Education				1.00	.207
r^2				.	.043
p (2-tail)				.	.003
N				1104	205
5 Exam					1.000
r^2					.
p (2-tail)					.
N					219

Notes:

[a] In a "pairwise" selection, sample size varies from pair to pair.

[b] In U.S. discrimination jurisprudence, statistical significance is generally accepted when $p < .05$ (similar to the social sciences). With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .005 (.05/10).

[c] p (1-tail).

The correlations of particular interest were those between sex and age (.084), education (.175), and Police Intake Exam score (.161), which, although slight to low, were similar to that between OAR and sex (.121). As with OAR and sex, all correlations

were significant,[21] which confirmed the patterns found in Tables 5-8, 5-9, and 5-10, indicating that female candidates were slightly older, better educated, and had higher Police Intake Exam scores (i.e., mental ability). These correlations will be explored in greater detail in the next section.

Because it was possible to estimate OAR reliability ($R \cong .60$), correlations between OAR and sex, age, education, and Police Intake Exam were corrected for attenuation. As indicated in Table 6-10 (which shows pairwise correlations corrected for attenuation), the correlation between OAR and sex was corrected upward from .121 to .156 (a difference of .035)[22] and the coefficient of determination was corrected upward from .015 to .024 (a difference of .009), and correlations and coefficients of determination between OAR and age, education, and Police Intake Exam score were similarly corrected.

Table 6-10

Pairwise correlations (1978-1999) corrected for attenuation

	Sex	Age	Educ.	Exam
OAR[a]	.156	.227	.132	.287
r^2	.024	.052	.017	.082
p (2-tail)[b]	.000[c]	.000	.000	.000
N	2941	1166	1104	219

Notes:

[a] OAR has an estimated corrected reliability of .60 ($R \cong .60$).

[b] In U.S. discrimination jurisprudence, statistical significance is generally accepted when $p < .05$ (similar to the social sciences). With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0125 (.05/4).

[c] p (1-tail).

[21] With a Bonferroni adjustment, the correlation between sex and Police Intake Exam was not significant; however, the sample size was only 219.

[22] A Fisher's Z transformation is unnecessary when coefficients are small.

These results highlight the fact that unreliability in testing (and measurement error in general) may conceal potential discrimination, which is of great practical value for proactively addressing discrimination,[23] although it may be of little probative value because the correction is theoretical. Nevertheless, although focussing on the contentious use of proxy variables rather than test scores, Paetzold and Willborn (1999) note that a failure to find discrimination in employment cases may be due to the problem of attenuation (§ 6.12).

Partial correlations r_p. The results of the bivariate correlations indicated that sex, age, education and Police Intake Exam score were all similarly related to OAR. However, at the aggregate level the fact remains that OAR was significantly related to sex (i.e., OAR was, in part, a function of sex), which if left unexplained may lead to an inference of discrimination. Therefore, to estimate the correlation between OAR and sex as if all candidates were equal in terms of age, education, and Police Intake Exam score, partial correlations were calculated to remove the confounding effects of these variables. Information regarding age, education, and Police Intake Exam score was limited to samples of candidates who had been hired; but as argued in Chapter 5, these samples were appropriate for analysis (the proportion of males and females and their average scores were similar to the population). However, information regarding age, education and Police Intake Exam scores was not equally available, and so it was necessary to compare the respective pairwise samples to determine if significant differences existed between them.

[23] The importance of correcting correlations for unreliability to prevent discrimination proactively will be addressed again later in this chapter and in Chapter 7.

Table 6-11 displays the comparison results, which indicate that the sample for which information was available on OAR, sex and age was almost identical to that for OAR, sex, age and education, where the difference in means was only .06 (38.41-38.35). The sample for which information was available on OAR, sex, age, education and Police Intake Exam scores, as discussed in Chapter 5, was a systematic random sample of Vancouver Police candidates who had attended the Police Academy assessment center and were subsequently hired. Although sample size was relatively small (N = 205), the mean was still quite similar to the other two samples, where the maximum difference was only .38 (38.41-38.03).

Table 6-11

Group comparisons (1978-1999): Listwise

	Total Candidates		Candidates Hired		
	OAR	OAR Sex	OAR Sex, Age	OAR Sex, Age, Educ.	OAR Sex, Age, Educ., Exam
N	2946	2941	1166	1101	205
Mn	36.77	36.78	38.35	38.41	38.03
Md	38.00	38.00	38.00	38.00	38.00
Sd	4.26	4.26	3.42	3.32	3.21
Min	15	15	22	25	28
Max	48	48	48	48	45

To determine if any of the differences between the sample means (reported in Table 6-11 above) were statistically significant, an analysis of variance (ANOVA) was calculated and the results are displayed in Table 6-12 below. These results indicate no significant differences ($p = .15$) between the samples of interest, which suggest that comparisons between pairwise partial correlations can be made (SPSS® Base 10.0, 1999, p. 183).

Table 6-12

ANOVA (one-way):[a] Comparing hired (listwise) groups[b]

Variance Source	SS	*df*	MS	F	*p*
Between conditions	38.3	2	19.15	1.74	.15
Within conditions	27159.0	2469	11.00		
Total	27197.3	2471	11.01		

Notes:

[a] ANOVA was calculated by hand, using the computation procedure described in Kachigan (1986, pp. 276-282). As the totals in each group were unequal and the short computational procedure was used, *n* for each group was replaced with an "harmonic" mean, which gives a conservative estimate of *F* (Rosenthal & Rosnow, 1991, pp. 318-319).

[b] From Table 6-11.

Table 6-13 below is a detailed correlation matrix that includes corrections for unreliability in the OAR and coefficients of determination. This table shows the results of the bivariate correlations between OAR and sex for each sample (listwise), along with the results of the partial correlations (pairwise), controlling for (1) age, (2) education, (3) age and education, (4) Police Intake Exam scores, (5) Police Intake Exam scores and age, (6) Police Intake Exam scores and education, and (7) Police Intake Exam scores, age and education. Notably, the results indicate that the size of the correlation coefficient between OAR and sex systematically decreases as the sample size decreases. Although this may be explained in part by an increasing margin of error for smaller samples, a relationship between the coefficient (between OAR and sex) and sample size is apparent (unadjusted $r = .77$), despite the fact that ANOVA indicated no significant differences between sample means (rf. Table 6-12). Consequently, any comparisons between pairwise partial correlations must be made cautiously.

Despite this limitation, an examination of Table 6-13 indicates that when the effects of age, education, and Police Intake Exam score were removed, separately and collectively, correlations between OAR and sex were reduced.

Table 6-13

Listwise bivariate (OAR with sex) and pairwise partial correlations (1978-1999): Corrected[a] and uncorrected

B: Bivariate P: Partial	df[b] df[b]	r r_p	r^2 r_p^2	B: Bivariate P: Partial	df[b] df[b]	r r_p	r^2 r_p^2
B: OAR – Sex[c]	2941	.121* [+]	.015	B: OAR – Sex[c]	217	.093	.009
Corrected		.156* [+]	.024	Corrected		.120*	.014
P: n/a	n/a	n/a		P: Exam[d]	216	.059	.003
				Corrected		.076	.006
B: OAR – Sex[c]	1164	.112* [+]	.013	B: OAR – Sex[c]	216	.091	.008
Corrected		.145* [+]	.021	Corrected		.118*	.014
P: Age[d]	1163	.099* [+]	.010	P: Exam/Age[e]	214	.057	.003
Corrected		.128* [+]	.016	Corrected		.074	.006
B: OAR – Sex[c]	1102	.110* [+]	.012	B: OAR – Sex[c]	203	.064	.004
Corrected		.142* [+]	.020	Corrected		.083	.007
P: Education[d]	1101	.094* [+]	.009	P: Exam/Educ[e]	201	-.011	.000
Corrected		.121* [+]	.015	Corrected		-.014	.000
B: OAR – Sex[c]	1099	.109* [+]	.012	B: OAR – Sex[c]	203	.064	.004
Corrected		.141* [+]	.020	Corrected		.083	.007
P: Age/Educ[e]	1097	.074*	.005	P: Exm/Age/Ed[f]	200	-.022	.001
Corrected		.100* [+]	.010	Corrected		-.028	.001

Notes:

[a] Corrected for attenuation, where OAR has an estimated corrected reliability of .60 ($R \cong .60$).

[b] df = N-2-order.

[c] Zero-order correlation.

[d] First-order partial correlation.

[e] Second-order partial correlation.

[f] Third-order partial correlation.

n/a indicates not applicable.

* Statistically significant ($p < .05$, one-tailed). In U.S. discrimination jurisprudence, statistical significance is generally accepted when $p < .05$ (similar to the social sciences).

[+] Statistically significant with a Bonferroni adjustment ($p < .006$, one-tailed), where the significance level (.05) is divided by the total number or correlations (8).

By noting the differences[24] between the bivariate correlations (OAR and sex) and the partial correlations for each configuration, the variables most responsible for reducing the correlation between OAR and sex were easily identified.[25] These variables were age and education collectively (difference = .035), the Police Intake Exam separately (difference = .034), and the Police Intake Exam and education collectively (difference =

[24] Fisher's Z transformation is unnecessary when correlation coefficients are small.

[25] Differences between corrected coefficients were not calculated because their utility lies generally in prevention rather than probity.

.064).[26] Notably, age and years of education, by themselves, were relatively unimportant variables for explaining the correlation between sex and OAR, a finding which was consistent with the fact that age and education differences between males and females (less than one year for both) did not appear substantive (rf. Tables 5-8 and 5-9 respectively). However, even though the difference between male and female scores was just over 3% (rf. Table 5-10), the Police Intake Exam, by itself, was a relatively important variable for explaining the correlation between sex and OAR. Similarly, when controlled collectively, the Police Intake Exam and education variables were relatively important variables. Specifically, when the effects of Police Intake Exam scores and education on OAR were removed, the correlation between candidate sex and OAR, although weak (.064), was reduced to zero. In other words, when males and females scored similarly on the Police Intake Exam and were of similar education, the relationship between candidate sex and OAR was essentially eliminated.

Binomial Effect-Size Display

Notwithstanding the results of the statistical tests, the binomial effect-size display (BESD), shown in Table 6-14 below, is a convenient way to interpret the practical significance of a correlation coefficient (*r*) (Rosnow and Rosenthal, 1996, p. 257). Based on the original aggregate data (rf. Table 6-6), Table 6-14 indicates that 12.1% more female candidates received passing scores than did male candidates (56.05 – 43.95).[27]

---

[26] The arithmetic difference is over .064 (.064 – (-.011)) = .075, but in these circumstances, where the discriminatory effect against males is the issue, it makes no sense to include a negative correlation coefficient (which indicates a main effect in favour of males).

[27] The results of the BESD (indicating an adverse effect of 12.1% against male candidates) are very similar to those of the parity analysis (indicating an adverse effect of 13.2%) (see Table 6-7). The BESD, then, is a very quick and useful method for determining the practical significance of a correlation coefficient.

Table 6-14

BESD:[a] Relationship of OAR (pass/fail)[b] with sex (population data)

A: M/F (1978-1999): pass/fail totals

Results	Male	Female	Total
Fail	1046	138	1184
Pass	1422	335	1757
Total	2468	473	2941

B: BESD (1978-1999): $r = .121$[c]

Results	Male	Female	Total
Fail	56.05[d]	43.95[e]	100
Pass	43.95[e]	56.05[d]	100
Total	100	100	200

Notes:
[a] Rf. Rosnow and Rosenthal (1996, pp. 257-263).
[b] OAR was transformed into a dichotomous variable (pass/fail) by splitting it at the median score (i.e., 38).
[c] 15.6 when corrected for attenuation.
[d] Calculated from: $[100(.5 + r/2)]$.
[e] Calculated from: $[100(.5 - r/2)]$.

When adjusted for unreliability in OAR scores, the coefficient of .121 was corrected to .156, which means that if reliability were increased, then the disparity between males and females would be greater yet. Because this correction is theoretical, its probative usefulness is likely limited when applied to a specific case. However, the practical implication for an assessment center is that if OAR is in fact a function of sex to some degree, then the correction demonstrates the consequences in terms of increased adverse effect should reliability be increased. Because high reliability is a desired test property, this relationship may at first appear paradoxical; however, if OAR is not a function of sex, then increasing reliability will not increase adverse effect.

Returning to the observed coefficient of .121, in practical terms this means that approximately 12.1% or 299 more male candidates (.121 x 2468 = 299) would have passed had they been female candidates.[28] In other words, 1721 males (compared to 1422) would have passed, and the pass rate would have been 70% (1721 ÷ 2468), similar

[28] The BESD is also useful for putting the coefficient of determination (r^2) into a practical perspective. For example, .121 squared equals .015, which technically suggests that only 1.5% of the variability in OAR is explained by sex, while the BESD demonstrates that the practical effect is 12.1%.

to the 70.8% pass rate of females (335 ÷ 473). Therefore, despite the inference of no discrimination according to the four-fifths rule, the adverse effect against 299 male candidates appears to be practically significant, which was also the case statistically, and so is prima facie legally significant.

A legal inference of discrimination, however, may be erroneous. It is based on the aggregate data and does not properly consider the "qualified" candidate, as previously discussed. For example, when the variables of age and education were collectively controlled (see Table 6-13), the bivariate correlation was reduced by .035 (i.e., from .109, which was statistically significant, to .074, which was not statistically significant with a Bonferroni adjustment). Here, in terms of a BESD, the adverse effect against male candidates was reduced from 10.9% (269) to 7.4% (182). When the variables of Police Intake Exam and education were collectively controlled, the bivariate correlation was reduced from .064 to -.011 (essentially zero); or in terms of a BESD, the adverse effect against males was reduced from 6.4% (158) to 0% (0). Therefore, assuming that age (i.e., maturity), education, and Police Intake Exam score (i.e., mental ability) are legitimate job prerequisites for the policing profession, a legal inference of discrimination on the basis of sex is probably not supported by the quantitative evidence.

Summary and Conclusions

With respect to systemic sex discrimination at the Police Academy assessment center, taking a quantitative approach this chapter reported and discussed the results of the "discriminatory effect" assessment phase. First, an analysis of parity indicated differences in OARs that were sex based, and so an analysis of probity was conducted to

determine if the differences were legally significant. The probity analysis was based on the four-fifths rule, statistical tests, and a test of practical significance.

The four-fifths rule, which measures selection or pass rates against an absolute standard, indicated that the differences in scores were not legally significant. On the other hand, taking a univariate approach the statistical tests indicated that the differences in scores were probably legally significant. However, a multivariate approach indicated that the differences were probably not legally significant when the individual effects of the Police Intake Exam score were controlled, or when the collective effects of age and education or education and Police Intake Exam score were controlled. The analysis of practical significance (BESD), when applied to the univariate approach indicated that the results were practically significant, but not when applied to the multivariate approach.

Statistically, the evidence suggests that the differences in scores were not primarily a function of sex but rather, in part, the effects of intervening variables, most notably age and education collectively, and education and Police Intake Exam scores collectively. As a result, there appears to be insufficient evidence to reject the null hypothesis of no discrimination. Legally, the evidence may also be insufficient to reject the null hypothesis of no discrimination, but only if the intervening variables are found to be material, which is a question of law, not fact.[29] This question, which is an unsettled area in law, highlights the limitations of a statistical analysis for assessing discrimination within a legal framework, which is discussed in greater detail in the next section. To

[29] Importantly, the inference that the Police Academy assessment center does not discriminate on the basis of sex is limited to the example presented in this study and so cannot be generalized to police recruitment, selection, and promotion in general.

conclude the discussion, the last section will briefly comment on the legend that has female candidates outperforming male candidates in assessment centers.

The Limitations of a Statistical Analysis

Ironically, one of the most legally contentious issues raised in this analysis is that of controlling for the effects of intervening variables,[30] such as age, education, and mental ability exam scores, because it effectively stratifies the aggregate data (i.e., applicant pool). In North America (including Canada and the U.S., except for cases involving compensation), existing case law suggests that the burden on plaintiffs is only to demonstrate discriminatory effect at the aggregate level; for example, that distinctions were made on the basis of sex (e.g., see Public Service, 1999).[31]

The problem with this approach is that it is univariate, and so can lead to erroneous conclusions both in terms of proving or disproving discrimination. For example, at the aggregate level primary results may indicate that the selection procedure (e.g., assessment center) is discriminating against males; however, if a relatively large proportion of females is better educated, stratification may indicate that the selection procedure is in fact differentiating on the basis of education, not discriminating on the basis of sex.[32] Alternatively, and of equal importance, is that aggregation may conceal discrimination, which stratification may reveal.[33] Therefore, stratifying aggregate data

[30] Note that, unlike its U.S. counterpart, the Supreme Court of Canada has not specifically considered the probative of statistical testing, although such evidence has been considered by Canadian tribunals, most notably Blake v. The Ministry of Correctional Services (1984). According to Vining et al. (1986), most studies examine disparities without controlling for other relevant factors (p. 661).

[31] See Vining et al. (1986, p. 683) for a Canadian commentary, and Paetzold and Willborn (1999, §§ 5.08, 5.09) for an American commentary. According to Paetzold and Willborn, regression analysis did not appear in U.S. compensation discrimination cases until 1975 (§ 6.01).

[32] In the context of studying discrimination in assessment centers, this was noted in the early 1970s. For example, MacKinnon (1975a) argued that comparing groups that had not been properly matched was methodologically unsound (p. 21).

[33] For a discussion and specific example, see Paetzold and Willborn (1999, § 5.08).

for analysis (by univariate statistical techniques or, equivalently, controlling for intervening variables by multivariate techniques) may lead to more precise conclusions about the causes and nature of discrimination.[34]

In law, however, although intervening variables such as education and mental ability exam scores may explain a discriminatory effect, this may be insufficient to negate an inference of discrimination. Unless it can be shown that such qualifications are not only business necessities but also meet the Supreme Court's three-step test for a BFOR, they will themselves be found discriminatory (rf. Public Service, 1999). Ostensibly neutral variables that directly or indirectly contribute to discrimination are considered "tainted," which highlights a limitation of statistical testing; i.e., it is the law, not statistics, that identifies tainted variables.[35] Notwithstanding, Paetzold and Willborn (1999) argue that if identified, tainted variables should not be automatically removed from a statistical analysis as they may be strongly related to other variables, which may assist in making inferences about the causes of discrimination (§ 6.13).

Within discrimination litigation, another limitation of a statistical analysis is that relating to the concept of the null hypothesis. Within a traditional statistical framework,[36] the null hypothesis is necessarily that of no discrimination, and so the alternate hypothesis is that of discrimination. For example, although this study identified females

---

[34] Stratifying data for a univariate test (e.g., a binomial distribution) is sometimes more limited than a multivariate analysis because it requires that the variables be defined categorically, but it also may be more intuitively accessible to those not familiar with statistical testing. In any case, the purpose of these analyses is the same, which is to break down aggregate data in an effort to obtain additional information that may be relevant to an inference about discrimination. For more discussion on this issue, see Baldus and Cole (1980), Connolly and Peterson (1980), Paetzold and Willborn (1999), Vining et al. (1986), and Zeisel and Kaye (1997).

[35] For example, see the classic U.S. case of Griggs v. Duke Power Co (1971), frequently cited by the Supreme Court of Canada, where education was held not to be a BFOR (and in fact contributed to discrimination) and so was not material to a prima facie case of discrimination.

[36] As distinguished from a Bayesian analysis, discussed in Chapter 5.

as the possible object of discrimination, the null hypothesis for this study is necessarily that of no discrimination. The limitation, then, is that a traditional statistical analysis cannot estimate the probability that discrimination has occurred;[37] rather, it estimates the probability of finding the observed results given the null hypothesis. If the predetermined alpha level is reached, it may lead to a rejection of the null hypothesis, but to conclude that discrimination has occurred requires an inferential leap by the fact-finder. The risk, of course, is making a Type I error, which is to reject the null hypothesis of no discrimination when it is in fact true.

Other limitations of a statistical analysis include the following: (1) the possibility of unidentified confounding factors that can invalidate inferences; (2) statistical significance as a function of the "power" of the test (which is in part a function of sample size); (3) primary data distorted by transformations; (4) measurement error (i.e., unreliability); (5) meeting underlying assumptions (e.g., a normal distribution, etc.); and (6) multicollinearity.[38] In short, a fundamental limitation of a statistical analysis is that it cannot conclusively prove or disprove discrimination; it can only provide circumstantial evidence, the probative weight of which is necessarily left to the courts.

The Legend of Females Outperforming Males

The results of this study are consistent with those of previous studies that find a main effect in favor of females at assessment centers (Walsh, Weinberg & Fairfield, 1987; Friedman, 1984; Gaugler et al., 1987; Hamner, Kim, Baird & Bigoness, 1974;

[37] This logic may seem counterintuitive to discrimination cases, where the probability of discrimination is the real issue of interest (Paetzold & Willborn, 1999, §§ 2.04, 4.13).

[38] Multicollinearity refers to variables that are linearly related to each other, which is a concern in correlational or regression analysis because if such variables cannot be isolated, then conclusions about individual predictor variables are tentative at best.

Mobley, 1982; and Peters, O'Conner, Weekley, Pooyan, Frank & Erenkantz, 1984).[39] But, where previous studies have not been able to offer a satisfactory empirical explanation for their results, generally or specifically (i.e., for particular occupations), this study suggests that for police the answer may lie partly in the fact that female candidates are on average older (possibly more mature),[40] better educated,[41] and have greater mental ability (as measured by the Police Intake Exam).

Of particular importance is the finding that mental ability appears to be a significant intervening variable for removing the effects of sex on OAR at the assessment center. It seems self evident that some level of intelligence or education is necessary to perform effectively within any occupation, and empirical evidence has shown that mental ability tests, such as intelligence tests and general aptitude tests, have high predictive validity for those occupations requiring complex cognitive skills (Ree & Earles, 1992). Therefore, if an assessment center that tests for similar occupations is valid,[42] then the hypothesis that assessment center performance is in part a function of mental ability merits attention (Klimoski & Brickner, 1987, p. 251).[43]

Since Bray's (1964) Management Progress Study (MPS), which utilized a number of mental ability tests, assessment center literature has consistently recognized the importance of intelligence. In 1973 Huck found a correlation of .40 between assessment

---

[39] Also discussed in the sections on discrimination in Chapter 2.

[40] In a meta-analysis of assessment center validity, Gaugler et al. (1987) reported that age was not a moderator of predictive validity.

[41] Huck (1973) found that education by itself did not explain assessment center scores, which is not inconsistent with the results of this study. When education was separately controlled, the effect of sex on OAR was minimally reduced.

[42] Regarding mental ability as a predictor at managerial assessment centers, see Howard (1997, p. 19) and Klimoski and Brickner (1987, pp. 251-252); and as a predictor at police assessment centers, see Gavin and Hamilton (1975, p. 172).

[43] For example, the Police Academy assessment center claims to measure job-related dimensions such as oral and written communications abilities, practical intelligence, and the ability to learn.

center OARs and School and College Ability Test (SCAT) scores, which has generally been confirmed by subsequent research (for example, see Byham, 1980a;[44] Hoffman & Thornton, 1997; Klimoski & Brickner, 1987; Moses, 1973; and Tziner & Dolan, 1982).[45] With respect to sex and mental ability, Howard and Bray (1988) found no overall difference between males and females, although females scored "notably" higher on the Verbal scale of SCAT, which was used as a measure of mental ability (p. 289).

This, then, begs the question of why females systematically scored higher than males on the Police Intake Exam (see Table 5-10),[46] which, when controlled, removed the effects of sex on OAR. If, as noted above, females do not generally have more mental ability than males, an alternative explanation is that females who have more mental ability (and who are older and better educated) apply for the policing profession. For example, the sample for this study was not representative of females generally, but rather represented those females who applied for a position as a municipal police officer.

In conclusion, the results of this study bring into question some of the hypotheses that have been advanced in assessment center literature. For example, one hypothesis suggested that because assessment center dimensions are often, by definition, dependent on interpersonal skills,[47] candidates who are more social, sympathetic and sensitive have an advantage. Therefore, if females are more caring police officers than males, as

---

[44] Byham (1980a), from a 1971 report, while discussing the research of Moses (1971), stated that "the [predictive validity] correlation of .461 was raised only minutely with the use of a mental ability test" (p. 32; rf. Moses, 1973).

[45] See Adams and Thornton (1988) for a review of the literature.

[46] A *t*-test comparing independent means indicated that the overall difference was significant, with p (t_{217} = -3.29) < .05 (two-tailed, equal variances assumed).

[47] Of particular relevance is that assessment centers generally test for "professional" occupations (such as business management, teaching, and policing), which require candidates to display high interpersonal skills (discussed in Chapter 5).

suggested by Balzer (1976), then they may have an advantage at an assessment center (p. 127). This, however, is contrary to the general assessment center research of Ritchie and Moses (1983), who found that the management style of women in management assessment centers and on the job was no different than that of males.

This study also brings into question two other hypotheses, which also serve to illustrate the speculation that has occurred on this subject. First, Gaugler et al. (1987) suggested that females may make better assessment center candidates because they are more self-disclosing than males (p. 504, citing Fletcher, 1981; rf. also Fletcher & Spencer, 1984). Second Walsh et al. (1987) suggested that female candidates may be favored by male raters (i.e., exhibiting errors of leniency) because they are sympathetic to females who have persevered in gaining access to traditionally male dominated occupations.

Compared to previous research, then, the present study has advanced a more empirical explanation for why females often outperform males in assessment centers, although the present study was limited to police recruits and a nonrandom sample.

CHAPTER VII: SUMMARY AND RECOMMENDATIONS

With substantive equality as the concern, the goal of this study was to provide a model to assess and prevent systemic discrimination within an assessment center. To test the utility of the proposed model, it was applied operationally to the Police Academy assessment center to determine whether it discriminated on the prohibited ground of sex. As a result, it was possible to evaluate the model in an applied context, provide the Police Academy with important information on its practice, and contribute to the literature on personnel selection. The purpose of this chapter, then, is to summarize the results of the discrimination assessment, evaluate the proposed model, and provide policy recommendations to the Police Academy assessment center.

The Assessment of Discrimination

Because the overall concern of this study was that of substantive equality as defined by law, this study was necessarily framed by an interpretation of discrimination that is operationally defined in terms of discriminatory effects, which necessarily involves comparisons between individuals or groups. And because systemic discrimination is often revealed by analyzing aggregate or group effects, the model proposed in this study incorporated a quantitative approach that includes the use of statistical tests. However, in Canada the probative value of statistical tests in discrimination litigation is not well defined, so, for guidance, this study turned to the U.S., where statistical tests have played an important role in discrimination litigation.

Subsequently, the proposed model, which is summarized in Figure 7-1, was applied to the Police Academy assessment center.

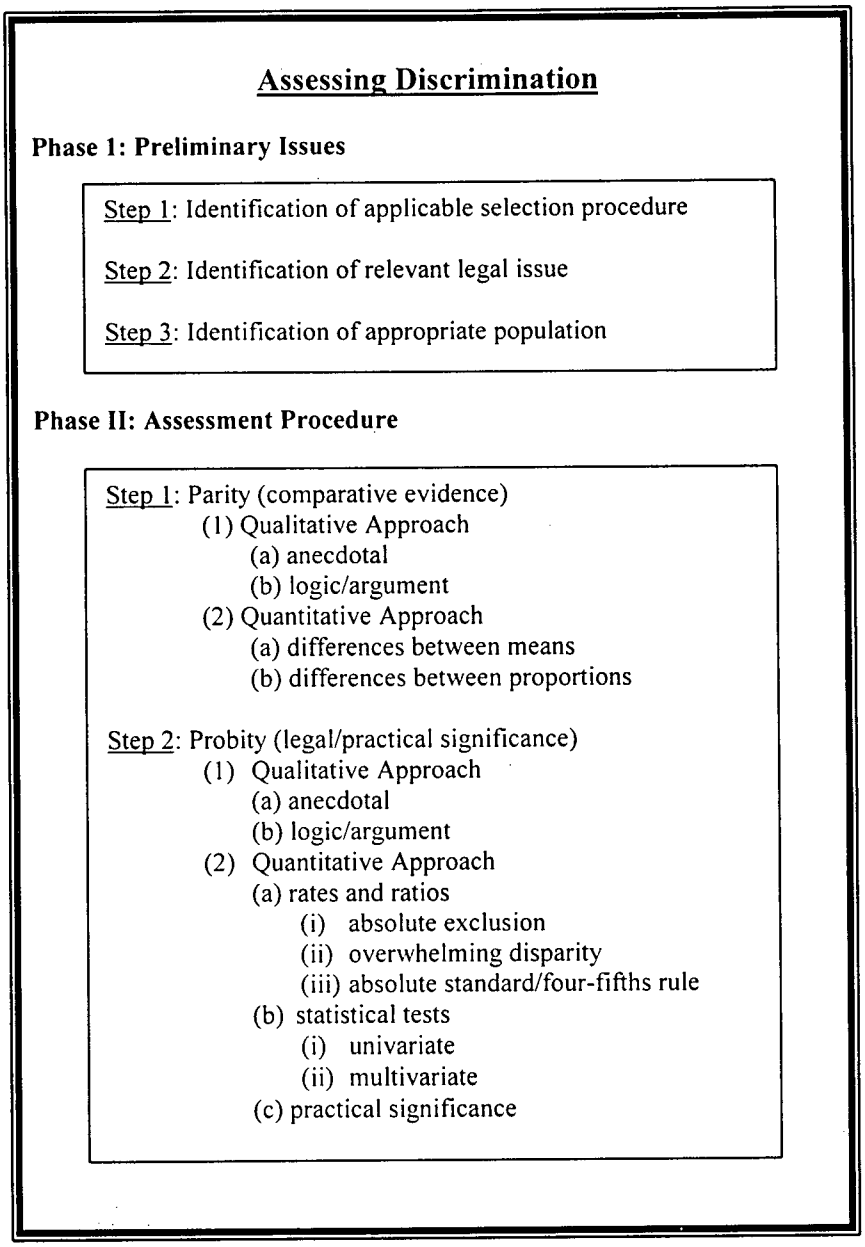


Figure 7-1. Model to assess discrimination[1]

[1] This study did not specifically discuss "absolute exclusion" and "overwhelming disparity." These standards are relatively straightforward and for the purposes of this study require little explanation. Absolute exclusion, as described in Lasani v. Ontario (Ministry of Community and Social Services) (1993), is a "ratio of non-representation," where no members of a "minority" group are present. Overwhelming disparity is somewhere between absolute exclusion and "disproportionately negative," but is still relatively easy to identify. It is when the disparity is less than overwhelming that fact-finders and lawmakers encounter considerable difficulty in identifying discrimination. In the U.S., in an attempt to address this difficulty, "absolute standards," such as the four-fifths rule, have been introduced (rf. Vining, McPhillips & Boardman, 1986, pp. 679-680).

In Phase I, which includes three steps, the preliminary issues were identified. In Step 1, the applicable selection procedure was identified; i.e., the Police Academy recruit assessment center. For this Step, Chapter 2 described the history and theoretical underpinnings of the assessment center method in general, and Chapters 1 and 5 described the Police Academy assessment center in particular.[2] In Step 2, the relevant legal issue was identified; i.e., discrimination (systemic) on the prohibited ground of sex. For this Step, Chapter 4, starting with the idea of justice and fairness, described the evolution of discrimination in Canadian law in order to frame this study and to theoretically ground the operational definition of systemic discrimination. In Step 3, the appropriate population was identified; i.e., all candidates who attended the assessment center between 1978 and 1999, where the groups of interest were males and females. For this Step, Chapter 5 described in detail the samples and population, along with the data collection techniques.

In Phase II, which includes two steps, the discrimination assessment procedure was conducted. In Step 1, taking a quantitative approach, parity was assessed; i.e., the scores of males were compared to those of females to determine if systematic differences existed between them. For this Step, Chapter 5 described the assessment procedures, while Chapter 6 reported and discussed the results, which were reported under two separate categories (means and proportions). Here, when mean scores and rating distributions were compared, the results indicated that females consistently received

[2] For Steps 1, 2, and 3 of Phase I, such a detailed analysis is unnecessary for a proactive assessment (discussed in the next section), although it may be necessary for litigation purposes.

higher scores than males. In other words, there appeared to be prima facie evidence of systemic discrimination on the prohibited ground of sex.

In Step 2, again taking a quantitative approach, probity was assessed; i.e., the differences found in Step 1 were analyzed to determine if they were legally (or practically) significant. Again, Chapter 5 described the assessment procedures, while Chapter 6 reported and discussed the results, which were reported under three separate categories. In the "rates and ratios" category, the differences between the scores (translated into pass-fail categories) of males and females were not legally significant according to the four-fifths rule (which requires that the pass rate of the sub-group exceeds .8 of the pass rate of the dominant group), where the adverse effect rate of 81.4% for males exceeded the arbitrary threshold of 80%.[3] However, as the sample sizes were large (greater than 200) and the adverse effect rate was close to the threshold, the evidence here was inconclusive.[4]

In the "statistical tests" category, the results of the univariate tests, which included a *t*-test comparing independent means (males with females) and a bivariate correlation *r* (OAR with sex), indicated that the scores of females were significantly higher than those of males ($p < .000$). Subsequently, in the "practical significance" category, the correlation between OAR and sex (.121) was assessed by using the binomial effect-size display technique. The results indicated that the correlation was practically significant,

[3] Under the "rates and ratios" category, an analysis of "absolute exclusion" was not relevant and an analysis of "overwhelming disparity" was not necessary.

[4] This conclusion is based on how the four-fifths rule is applied in the U.S., which was discussed in Chapters 1 and 4 (rf. Paetzold & Willborn, 1999, §§ 5.06, 5.07). Although the four-fifths rule has not knowingly been applied in Canada, it may well have some probative value because in discrimination cases Canadian courts often look to the U.S. for guidance (Vining et al., 1986, p. 662).

where 299 out of 2,468 (12.1%) male candidates were adversely affected.[5] Therefore, at the aggregate level the compendium of evidence (adverse effect rate and significant differences, both statistically and practically) indicated that the differences between the scores of females and males might be sufficient to constitute legal significance.

However, a more detailed analysis of the aggregate data provided evidence to the contrary. First, in an exploratory analysis, descriptive information reported in Chapter 5 indicated that females were slightly older, better educated and scored higher on the Police Intake Exam.[6] That females have been somewhat older and better educated has been noted in the literature in both Canada (Moore, 1997, p. 39) and the United States (Miller, 1998, p. 162),[7] but a satisfactory empirical explanation has not been advanced and is beyond the scope of this study. Notably, the differences between males and females on these variables did not appear substantive (both are less than one year), but the bivariate correlations reported in Chapter 6 indicated that age, education, and Police Intake Exam, along with sex, were significantly related to OAR.[8] This information suggests that despite the results of the univariate tests, the difference in the scores of males and females may be better explained by variables other than sex. Next, in a relatively simple

---

[5] Practical significance does not necessarily follow statistical significance and statistical significance does not necessarily follow practical significance (Paetzold & Willborn, 1999, §§ 4.11, 4.12, 4.13, 4.16, 6.07; Vining et al., 1986, p. 681). Practical significance will depend on the facts and circumstances of each case and in the final analysis is a matter of the adjudicator's judgment.

[6] In Chapters 5 and 6, the assumption was made that the Police Intake Exam measured mental ability.

[7] Data reported by Dantzker and Kubin (1998) do not support this phenomenon, although their sample was purposive (p. 25).

[8] Despite being ostensibly neutral, subtle patterns such as this may reveal long-standing institutional discrimination. For example, Moore (1998) speculates that Canadian police departments were looking for more "mature" females, which she suggests may provide an explanation for police departments hiring older and better educated female officers (p. 39). It may be, then, that municipal police departments in British Columbia have historically discriminated against females by only nominating their idea of "mature" females to attend the Police Academy assessment center. Notwithstanding, as previously discussed in Chapter 5, because the Police Academy plays no role in candidate selection, only the candidate pool is material in an assessment of the Police Academy.

multivariate analysis, the results of partial correlations r_p reported in Chapter 6 indicated that once the confounding effects on OAR by education and Police Intake Exam score collectively were removed, the listwise correlation between OAR and sex (.064) was reduced to zero.[9]

In other words, when male and female candidates were of similar education and scored similarly on the Police Intake Exam, the differences between their scores were not significant, either statistically or practically. Legally, however, the differences between scores may yet be significant because prima facie evidence of discrimination at the aggregate level, in these circumstances, is still material unless it can be shown that level of education and performance on the Police Intake Exam are bona fide occupational requirements for policing. If not, regardless of their explanatory value, these variables have little probative value in deciding the question of discrimination.

The different results obtained from the univariate and multivariate tests show how a statistical analysis, legally, can be inconclusive and in addition can raise contentious questions, such as whether aggregate data should be stratified. As illustrated here, depending on how the results are interpreted, it may be argued that controlling for the effects of intervening variables either explains discrimination or conceals it. Other basic questions that are unique to a statistical analysis include those relating to the framing of the null hypothesis, the use of one-sided or two-sided tests, and the meaning of significance (Paetzold & Willborn, 1999, §§ 4.14, 5.07, citing Baldus & Cole, 1980; see also Vining, McPhillips & Boardman, 1986).

[9] The results of the statistical tests were consistent with this study's hypotheses, which is of interest academically but not legally.

Moreover, some multivariate statistical models can be completely overwhelming and incomprehensible and so are not well suited for answering basic questions about discrimination (Baldus & Cole, 1980, pp. 6-7, 74-75, cited by Vizkelety, 1987, p. 175). Admittedly, sophisticated statistical techniques such as multiple regression have the means to construct derived (and complicated) variables that can be quite informative; but the more the results are removed from the primary data, the more likely are the chances for distortion, error, and disagreement. And such results, which are often ambiguous to statisticians and inaccessible to those who are not, are difficult to apply in the real world.[10]

Because the facts and circumstances that provide the necessary context for each discrimination assessment will shape a quantitative analysis, the model proposed in this study cannot recommend any particular statistical tests. But as recommended in Chapter 3, the principle underlying the metaphor of Occam's razor should be used to select the most direct and common tests so that the issues are not unnecessarily complicated beyond that which already exists in law. For example, the results of this study were based on simple differences between scores, a basic analysis of rates and ratios, and intuitively accessible correlations. As noted by Kachigan (1991), "The simplest analyses are often the best analyses" (p. 159; rf. also Rosenthal & Rosnow, 1991, pp. 32, 391).

[10] A judge of a federal court in the U.S. once lamented, "The courts have unintentionally opened a Pandora's Box by using the word 'statistics' instead of 'percentages' because Title VII cases [have become] contests between college professor statisticians who revel in discoursing about statistical theory" (cited in Vining et al., 1986, p. 684).

Interrater Reliability and Discrimination

Twenty years ago Shrout and Fleiss (1979) pointed out that judgments made by humans were plagued by problems of reliability and bias, which were especially evident in assessment centers (p. 420), and the situation is no different today. In response, the goal of this study, in addition to providing a model to assess discrimination, was to provide a means to prevent discrimination within an assessment center. Here, the analysis focuses more on measurement issues (i.e., interrater reliability) than on legal issues because the objective is to be proactive and thereby avoid unnecessary discrimination litigation.

In contrast to traditional pencil and paper tests, such as the commercially available Wonderlic Personnel Test (WPT), a generalized reliability coefficient cannot be calculated for the assessment center method because rater assessments cannot be standardized to the degree possible for item responses in a written test. Because interrater reliability is a function of a number of factors (such as training, aptitude, mental ability, concentration, etc.) that differ from assessment center to assessment center, it is imperative that each assessment center conducts a separate interrater reliability analysis (discussed in detail in Chapter 3).

Although discrimination can be assessed without information about reliability, if discrimination were indicated by a correlation between assessment scores and candidate sex,[11] the results are insufficient for determining the potential (i.e., theoretical extent) of the problem. A detailed reliability analysis can often be quite complex and beyond the

[11] The results of a *t*-test (or *F*, when the numerator $df = 1$) are easily are easily converted into a correlation *r* (Rosnow & Rosenthal, 1996, pp. 276, 291; Rosenthal & Rosnow, 1991, p. 323).

mandate of assessment center administrators, but by using the correlational approach, as demonstrated in Chapter 3, an administrator can quite easily calculate overall interrater reliability. Once reliability is known, a correlation between assessment center scores and candidate sex (or race, religion, etc.) can be corrected for attenuation.

Based on the assumption that a correlation coefficient cannot exceed the square root of the product of the reliability coefficients of the variables of interest (Carmines & Zeller, 1979, p. 34, citing Lord & Novick, 1968; Cronbach, 1984, p. 176; Kaplan & Saccuzzo, 1982, p. 90), it follows that the reliability of a measure defines the upper limit of its correlation with another. Therefore, if candidate sex were correlated with assessment center scores (which is evidence of discrimination) but these scores were unreliable, the observed correlation coefficient would underestimate the true relationship between sex and scores. For example, knowing that the Police Academy's interrater reliability coefficient was on average .6, it follows that the observed bivariate correlation coefficient of .121 underestimated the true correlation between candidate sex and assessment scores (discussed in Chapter 6).

By applying the formula provided by Carmines and Zeller (1979), the observed coefficient of .121 was corrected to .156, which in practical terms, when translated by the binomial effect-size display, means that rather than 12.1% the adverse effect is potentially 15.6%. However, it is important to note that the correction is theoretical and so is most likely useful for the purposes of prevention rather than litigation. For example, theoretically, if the reliability of the Police Academy were .3 and the correlation between assessment center scores and candidate sex .2, then the corrected coefficient would be .37, almost twice as large. This correction illustrates how information about reliability

can provide important information about the extent to which discrimination may potentially exist should reliability be increased, allowing an assessment center administrator to be more proactive than otherwise possible.

Recommendations for the Police Academy

It is worth noting that the purpose of this study was not about finding fault with the Police Academy assessment center; rather, similar to that of a policy analysis, it was about "what ought to be done, about making things better, not worse" (Wildavsky, 1979, p. 13).[12] With this in mind, based on the results of the assessment of both discrimination and interrater reliability, this study concludes with the following practical suggestions and recommendations, which together should assist the Police Academy in proactively addressing discrimination.

Discrimination and Interrater Reliability Assessments

The Police Academy should assess discrimination regularly in both recruitment and promotional selection procedures, using the model proposed by this study.[13] Moreover, consistent with this model, the Police Academy should assess interrater reliability regularly, using the correlational approach recommended by this study.[14]

---

[12] As noted by Wildavsky (1979), a good organization evaluates itself, its purpose being the recognition and correction of errors, which is the "essence of policy analysis" (pp. 16, 389, 393-394). For more discussion on the policy analysis process, see Haas and Springer (1998, p. 18) and Mood (1983, p. 6).

[13] Notably, the findings of such assessments would be limited to the Police Academy assessment center. To assess discrimination in police selection and promotion generally, a much broader analysis would be required (e.g., including the recruitment and selection procedures of the participating police departments in the analysis).

[14] As discussed in Chapter 2, research has shown that interrater reliability may be improved by ensuring that (1) proper attention is paid to rater selection, training and evaluation, (2) the exercises are relevant and the dimensions are clearly identified, (3) cognitive demands are realistic (e.g., limiting the number of complex dimensions), and (4) role playing is standardized (otherwise, the scores of candidates cannot be compared with each other).

Reliability provides important information about measurement consistency and the potential for discrimination.

Item Analysis and Internal Consistency Analysis

To improve upon the assessment scale, the Police Academy should periodically conduct an item analysis to assess the contribution of each dimension to the overall assessment and an internal consistency analysis to assess the inter-item consistency of the dimensions within the appropriate scales or sub-scales.[15]

Specific Dimension Recommendations

The Police Academy should review the dimension of flexibility, which was making a negligible contribution to the overall assessment and to internal consistency. The Police Academy should also consider classifying the dimension of written communications as a separate scale, as it appears to be measuring an attribute that is completely unrelated to the other dimensions.

Rater Evaluation

The Police Academy should ensure that raters are not certified unless they have objectively demonstrated performance competency at a predetermined level.[16] After certification raters should be periodically evaluated to ensure that their ratings continue to correlate well with those of other raters and to guard against common rater errors.[17]

---

[15] For research purposes, to assess the discriminant and convergent validity of the dimensions, the Police Academy may wish to consider recording how raters score dimensions within each exercise. Also, if the Police Academy were to ensure that at least two raters observed a candidate on each dimension, it would be possible to assess interrater reliability by dimension rather than just overall score.

[16] In England, where assessment centers are used extensively for police, the Objective Structured Performance Related Examination (OSPRE) center requires newly trained raters to pass a final examination. This examination requires newly trained raters to rate standardized scenarios that have already been rated by an expert group. If the ratings of the newly trained raters do not correlate well with those of the expert group, then they fail the course (Hutton & Sampson, 1999, p. 82).

[17] According to Byham, "It is important to note that cognitive training does not produce reliability. You only get that through practice and feedback against defined standards" (quoted in Mayes, 1997, p. 6).

Record Keeping

In addition to recording candidate names and scores, the Police Academy should record demographic information that can assist in an assessment of discrimination. In particular, for each assessment center class the Police Academy should record, at minimum, the age, sex, race, and education for both candidates and raters.[18]

Concluding Statement

It is important to note that although this study intended to provide an example of how a statistical analysis might be useful in assessing systemic discrimination, it did not intend to provide an analysis of statistical applications in discrimination litigation.[19] It is also important to note that while applying statistical models to real world situations (e.g., the Police Academy assessment center) raises a variety of issues, the main issues in any employment based analysis of discrimination are fundamentally legal. A statistical analysis may provide assistance to a court or tribunal in determining whether any differences between the groups of interest are practically significant, especially in cases of systemic discrimination, but in the final analysis the arguments are legal, not statistical, as are the relevant standards for proof (i.e., legal significance vis-a-vis statistical significance).[20]

---

[18] It is important to record information about raters because very little is known about how assessor characteristics effect scoring (Bartels & Doverspike, 1997; Lowry, 1993).

[19] For this, see Vining et al. (1986) and Vizkelety (1987), who provide a Canadian perspective, and Baldus and Cole (1980) and Paetzold and Willborn (1999), who provide an American perspective.

[20] A mathematical or strict logic argument is conceptually quite different from a legal argument, which is guided by different purposes, different rules of procedure, and is particularly dependent upon the facts and circumstances of each individual case. As noted by Mewett (1999), professor of Law Emeritus at the University of Toronto, "Proof, in law, is not a matter of mathematical logic, but of [legal] probabilities, dependent upon the standard of proof required. For another, if mathematical certainty were required, there would be no convictions and no successful plaintiffs" (p. 319).

Based on the results of the assessment of the Police Academy, the proposed model appears to provide a relatively uncomplicated and sensible approach, consistent with Canadian jurisprudence,[21] for assessing systemic discrimination in an assessment center. Moreover, because patterns of inequality are essentially the same regardless of the context, the proposed model is recommended for assessing discrimination in employment selection generally. Although the proposed model cannot define legal significance (i.e., whether the discriminatory effect was "disproportionately negative")[22] as a statistician defines statistical significance, it provides a means for its assessment insofar as is reasonably possible within the evolving definition of discrimination. Similar to discrimination law, the purpose of discrimination assessment should be prevention, and this study has also demonstrated how a reliability analysis can be an important tool for proactively addressing discrimination. It is clear from the issues raised in this study that assessing discrimination can raise difficult questions for which no easy answers exist, legal or otherwise; but the questions can be rationally argued, and it is to this end that this study has hopefully made a contribution.

[21] A review of Paetzold and Willborn (1999) suggests that the model is also consistent with American jurisprudence.

[22] The standard of a "disproportionately negative" effect has been discussed by the Supreme Court, for example, in Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987) and recently in Public Service (1999), but has never been defined.

REFERENCE LIST

Cases Cited

Action Travail des Femmes v. C.N. Railway (1987) C.H.R.R. D/4210 [When heard by S.C.C., indexed as Canadian National Railway Co. v. Canada (Canadian Human Rights Commission) (1987)].

Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

Alberta (Human Rights Commission) v. Central Alberta Dairy Pool, [1990] 2 S.C.R. 489.

Andrews v. Law Society of B.C., [1989] 1 S.C.R. 143.

Battlefords & District Co-operative v. Gibbs, [1996] 3 S.C.R. 566.

Berry, B., Stokes, E.L., Laut, K.E., v. City of Omaha and Wervel, L.H., [1975] No. 31 (Doc. 695), District Court of Douglas County, Nebraska.

Bhadauria v. Board of Governors of Seneca, [1981] 2 S.C.R. 181.

Bhinder v. C.N., [1985] 2 S.C.R. 561.

Blake v. The Ministry of Correctional Services and Mimico Correctional Institute (1984) 5 C.H.H.R. D/2417.

Blanchard v. Control Data Canada Ltd., [1984] 2 S.C.R. 476.

British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union (1997), unreported, Vancouver Registry CA22360 (B.C.C.A.) [appeal of arbitration award, 1996, 58 L.A.C. (4th) 160].

British Columbia (Public Service Employee Relations Commission) v. British Columbia Government and Service Employees' Union [1999] S.C.J. No. 46 (QL).

British Columbia Superintendent of Motor Vehicles and the Attorney General of British Columbia v. British Columbia Council of Human Rights (1997), unreported, Vancouver Registry V02837 (B.C.C.A.).

Brossard (Town) v. Quebec (Commission des droits de la personne), [1988] 2 S.C.R. 279.

Canada (Human Rights Commission) and Husband v. Canada (Armed Forces), [1994] 3 F.C. 188 (leave to appeal refused, (1994), 118 D.L.R. (4th) vi).

Canadian National Railway Co. v. Canada (Canadian Human Rights Commission), [1987] 1 S.C.R. 1114 [Cross reference to Action Travail des Femmes v. C.N. Railway (1987)].

Castaneda v. Partida, 430 U.S. 493 (1977).

Central Okanagan School District No. 23 v. Renaud, [1992] 2 S.C.R. 489.

Chambly, Com'n Scolaire v. Berguein (1994), 115 D.L.R. (4th) 609 (S.C.C.).

Deloitte, Haskins & Sells Ltd. v Alberta (Workers' Compensation Board), [1985] 1 S.C.R. 785.

Eaton v. Brant County Board of Education, [1997] 1 S.C.R. 241.

Edmonton Journal v. Alberta (Attorney General), [1989] 2 S.C.R. 1326.

Egan v. Canada, [1995] 2 S.C.R. 513.

Eldridge v. British Columbia (Attorney General), [1997] 3 S.C.R. 624.

Firefighters Institute v. City of St. Louis, [1977] 549 F. 2d 506 (8th Cir.).

Girvin v. The King, [1911] 45 S.C.R. 167.

Godbout v. Longueuil (City) (1997), 219 N.R. 1 (S.C.C.). (rf. Watt, D., & Fuerst, M. (1999). The 1999 annotated Tremeear's Criminal Code. Canada: Carswell Legal Publications, at 1332.)

Gould v. Yukon Order of Pioneers, [1996] 1 S.C.R. 353.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Hazelwood School Dist. v. United States, 433 U.S. 299 (1977).

Insurance Corporation of British Columbia v. Heerspink, [1982] 2 S.C.R. 145.

Large v. Stratford (City), [1995] 3 S.C.R. 733.

Lasani v. Ontario (Ministry of Community and Social Services) (1993), 21 C.H.R.R. D/415.

Little Sisters Book and Art Emporium v. Canada (Minister of Justice) (1998), 125 C.C.C. (3d) (S.C.C.) 484.

McDonnell-Douglas Corp. v. Green, 411 U.S. 792 (1973).

McKinney v. University of Guelph, [1990] 3 S.C.R. 906.

Martinueau v. Matsqui Institution Disciplinary Board, [1980] 1 S.C.R. 602.

Ontario Human Rights Commission v. Etobicoke, [1982] 2 S.C.R. 1297.

Ontario Human Rights Commission & O'Malley v. Simpson Sears Ltd., [1985] 2 S.C.R. 536.

P.S.A.C. v. Canada (Treasury Board) (1991), 14 C.H.R.R. D/341 (Can. Trib.).

R. v. Big M Drug Mart Ltd. (1985), 18 C.C.C. (3d) 385 (S.C.C.).

R. v. Oakes, [1986] 1 S.C.R. 103.

R. v. Sault Ste. Marie, [1978] 2 S.C.R. 1299.

R. v. Turpin, [1989] 1 S.C.R. 1296.

Reference re Language Rights under S. 23 of Manitoba Act, 1870 and S. 133 of Constitution Act, 1867, [1985] 1 S.C.R. 721.

Ross v. School District No. 15, [1996] 1 S.C.R. 825.

Saskatchewan (Human Rights Commission) v. Saskatoon [Firefighters], [1989] 2 S.C.R. 1297.

Saskatchewan (Human Rights Commission) v. Canadian Odeon Theatres Ltd. (1985), 18 D.L.R. (4th) 93 (Sask. C.A.).

Teamsters (International Brotherhood of) v. United States, 431 U.S. 324 (1977).

United States v. Test, 550 F. 2d 577 (10th Cir., 1976).

University of B.C. v. Berg, [1993] 2 S.C.R. 353.

Vriend v. Alberta, [1998] 1 S.C.R. 493.

Winnipeg School Division No. 1 v. Craton, [1985] 2 S.C.R. 150.

Zurich Insurance Ltd. v. Ontario (Human Rights Commission), [1992] 2 S.C.R. 321.

Statutes and Regulations Cited

Canadian Charter of Rights and Freedoms, Part I of the Constitution Act, 1982, being Schedule B to the Canada Act 1982 (U.K.), 1982, c. 11.

Canadian Human Rights Act, R.S.C. 1985, c. 4.

Constitution Act, 1967 (U.K.), 30 & 31 Vict., c. 3.

Constitution Act, 1982 being Schedule B to the Canada Act 1982 (U.K.), 1982, c. 11.

Court Order Enforcement Act, R.S.B.C. 1996, c. 78.

Human Rights Code, R.S.B.C. 1996, c. 210.

Individual's Rights Protection Act, 1996, S.A. 1996, c. 25.

Judicial Review Procedure Act, R.S.B.C. 1996, c. 241.

Police Act, S.B.C. 1996, c. 367.

Police Amendment Act, S.B.C. 1997, c. 37

Provincial Standards for Municipal Police Departments in British Columbia, O.I.C., 1992/748.

Rules Regarding Training, Certification and Registration of Municipal Constables Appointed Under Section [26] of the Police Act, B.C. Reg. 109/81.

Uniform Guidelines on Employee Selection Procedures (25 August, 1978). [United States] Federal Register, 43, 38290-38315.

Authors Cited

A pale, male reflection of the community: Equity still eludes Toronto police. (1999, March/April). Police Employment Law News. Canada: Lancaster's.

Abella, R. (1984). Report on the Commission on Equality in Employment. Ottawa: Minister of Supply and Services Canada.

Adair, M., & Moon, P. (1977). Managers for tomorrow: Identifying the future police executives. RCMP Gazette, 38 (9), 1-5.

Adams, S., & Thornton, G., III. (1988, May). The assessor judgment process: A review of the reliability and validity of assessment center ratings. Paper presented at the International Congress on the Assessment Center Method, Tampa, Florida.

Aiken, L. (1996). Rating scales and checklists: Evaluating behavior, personality, and attitudes. New York: John Wiley & Sons, Inc.

Alexander, H. (1947). The estimation of reliability when several traits are available. Psychometrika, 12, 79-99.

Baker, B., Hardyck, C., & Petrinovich, L. (1966). Weak measurement vs. strong statistics: An empirical critique of S.S. Stevens' proscriptions on statistics. Educational and Psychological Measurement, 26, 291-309.

Baldus, D., & Cole, J. (1980). Statistical proof of discrimination. CO: Shepard's Inc.

Balzer, A. (1976). A view of the quota system in the San Francisco Police Department. Journal of Police Science and Administration, 4 (2), 124-133.

Bartels, L., & Doverspike, D. (1997). Assessing the assessor: The relationship of assessor personality to leniency in assessment center ratings. Journal of Social Behavior and Personality [Special Edition], 12 (5), 179-190.

Bartko, J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

Bartko, J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Bender, J. (1973). What is "typical" of assessment centers? Personnel, 50, 50-57.

Berry, K., & Mielke, P. (1988). A generalization of Cohen's *kappa* agreement measure to interval measurement and multiple raters. Educational and Psychological Measurement, 48 (4), 921-933.

Berry, K., & Mielke, P. (1990). A generalized agreement measure. Educational and Psychological Measurement, 50, 123-125.

Bittner, L. (1980). The functions of the police in modern society. Cambridge: Olegeschlager, Gunn, & Hain.

Black, B. (1994). B.C. human rights review: A report on human rights in British Columbia. B.C.: Ministry Responsible for Multiculturalism and Human Rights.

Black, M. (1990). Black's law dictionary (6th ed.). St. Paul, Minnesota: West Publishing Co.

Bobrow, W., & Leonards, J. (1997). Development and validation of an assessment center during organizational change. Journal of Social Behavior and Personality [Special Edition], 12 (5), 217-236.

Borg, W., & Gall, M. (1983). Educational research: An introduction (4th ed.). New York: Longman.

Borman, W. (1982). Validity of behavioral assessment for predicting military recruits performance. Journal of Applied Psychology, 67, 3-9.

Bray, D. (1964). The management progress study. American Psychologist, 19, 419-429.

Bray, D. (1976). Identifying managerial talent in women. Atlantic Economic Review, 26, 38-43.

Bray, D. (1982). The assessment center and the study of lives. American Psychologist, 37 (2), 180-189.

Bray, D. (1999). Assessment centers: Centered on assessment [On-line]. Available: http://www.ddiworld.com.

Bray, D., & Campbell, R. (1968). Selection of salesmen by means of an assessment center. Journal of Applied Psychology, 52 (1), 36-41.

Bray, D., & Grant, D. (1966). The assessment center in the measurement of potential for business management. Psychological Monographs, 80 (17, Whole No. 625).

Bray, D., Campbell, R., & Grant, D. (1974). Formative years in business: A long-term study of managerial lives. New York: Wiley-Interscience.

Briscoe, D. (1997). Assessment Centers: Cross-cultural and cross-national issues. Journal of Social Behavior and Personality [Special Edition], 12 (5), 261-270.

British Columbia Police Commission. (1990). Provincial standards for municipal police departments in British Columbia. British Columbia: Author.

British Columbia Police Commission. (1995). Provincial standards for municipal police departments in British Columbia. British Columbia: Author.

British Columbia Police Commission. (1996). Review of the Justice Institute of British Columbia Police Academy. British Columbia: Author.

Brook, J. (1990). A lawyer's guide to probability and statistics. Ontario: Carswell Legal Publications.

Brown, G. (1978, June). What you always wanted to know about assessment centers but were afraid to ask. Police Chief, 60-67.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322.

Bruning, J. & Kintz, B. (1968). Computational handbook of statistics. Illinois: Scott, Foresman, & Co.

Burbeck, E., & Furnham, A. (1985). Police officer selection: A critical review of the literature. Journal of Police Science and Administration, 13 (1), 58-69.

Bureau of Justice Statistics. (1995, September). Law enforcement management and administrative statistics. Washington, DC: U.S. Department of Justice.

Bycio, P., Alvares, K., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. Journal of Applied Psychology, 72, 463-474.

Byham, W. (1970). Assessment Centers for spotting future managers. Harvard Business Review, 48 (4), 150-160.

Byham, W. (1977a). Application of the assessment center method. In J. Moses & W. Byham (Eds.), Applying the assessment center method (pp. 31-44). New York: Pergamon Press.

Byham, W. (1977b). Assessor selection and training. . In J. Moses & W. Byham (Eds.), Applying the assessment center method (pp. 89-126). New York: Pergamon Press.

Byham, W. (1980a). The assessment center as an aid in management development. Training and Development Journal, 24-36.

Byham, W. (1980b). Review of legal cases and opinion dealing with assessment centers and content validity. Pittsburgh: Development Dimensions International.

Byham, W. (1980c). Starting an assessment center the correct way. Personnel Administrator (February), 27-32.

Byham, W. (1999). What is an assessment center? [On-line]. Available: http://www.ddiworld.com.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carmines, E., & Zeller, R. (1979). Reliability and validity assessment (series no. 07-017). California: Sage.

Chamberlain, L. (1980). An initial study of the police constable selection program. British Columbia: Justice Institute of British Columbia Police Academy.

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

Cohen, S. (1980, December). Pre-packaged vs. tailor-made: The assessment center debate. Personnel Journal, 989-991.

Collins, J. (1985). Interrater reliability of assessors for the police academy assessment center. British Columbia: Justice Institute of British Columbia Police Academy.

Connolly, W., & Peterson, D. (1980). Use of statistics in equal opportunity litigation. New York: Law Journal Seminars-Press.

Coutts, L. (1990). Police hiring and promotion: Methods and outcomes. Canadian Police College Journal, 14, (2), 98-122.

Cramer, D. (1997). Basic statistics for social research. New York: Routledge.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16 (3), 297-334.

Cronbach. L. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row, Publishers.

Cronbach, L. (1984). Essentials of psychological testing (4th ed). New York: Harper & Row, Publishers.

Cronbach, L., Gleser, G., Nanada, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: John Wiley & Sons, Inc.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cronbach, L., Yalow, E., & Schaeffer, G. (1980). A mathematical structure for analyzing fairness in selection. Personnel Psychology, 33, 693-704.

Dantzker, M. (1994). Measuring job satisfaction in police departments and policy implications: An examination of a mid-size, southern police department. American Journal of Police, 13 (2), 77-101.

Dantzker, M., & Kubin, B. (1998). Job satisfaction: The gender perspective among police officers. American Journal of Criminal Justice, 23 (1), 19-31.

Davies, P. (Ed.). (1988). Human rights. London: Routledge.

Day, S., & Brodsky, G. (1996). The duty to accommodate: Who will benefit? Canadian Bar Review, 75, 433.

Deutsch, M. (1985). Distributive justice: A social-psychological perspective. New Haven: Yale University Press.

Dick, W., & Hagerty, N. (1971). Topics in measurement: Reliability and validity. New York: McGraw Hill.

Donahue, J., Truxillo, J., Cornwell, & Gerrity, M. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. Journal of Social Behavior and Personality [Special Edition], 12 (5), 85-108.

Dreher, G., & Sackett, P. (1981). Some problems with applying content validity evidence to assessment center procedures. Academy of Management Review, 6 (4), 551-560.

Duchesneau, J. (1997). Women in policing: An up-to-date account. In M. LeBeuf & J. McLean (Eds.), Workshop on women in policing (pp. 170-175). Ottawa: Canadian Police College.

DuPerron, W. (1997). Usefulness of the assessment centre approach to identifying management potential. Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta.

Ebel, R. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Ebel, R. (1972). Essentials of educational measurement. New Jersey: Prentice-Hall, Inc.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (25 August, 1978). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43, 38290-38315.

Erickson, B., & Nosanchuk, T. (1977). Understanding data. Toronto: McGraw-Hill Ryerson Ltd.

Eysenck, H. (1984). The effect of race on human abilities and mental test scores. In C. Brown & R. Brown (Eds.), Perspectives on bias in mental testing (pp. 249-292). New York: Plenum.

Felkenes, G., & Unsinger, P. (1992). Diversity, affirmative action and law enforcement. Illinois: Charles C. Thomas.

Feltham, R. (1988). Validity of a police assessment center: A 1-19 year follow-up. Journal of Occupational Psychology, 61 (2), 129-144.

Finn, R. (1970). A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 30, 71-76.

Fisher, R. (1932). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fitzgerald, L., & Quaintance, M. (1982). Survey of assessment center use in state and local government. Journal of Assessment Center Technology, 5, 9-21.

Flanagan, J. (1954). Some considerations in the development of situational tests. Personnel Psychology, 7, 461-464.

Fleenor, J., Fleenor, J., & Grossnickle, W. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. Journal of Business and Psychology, 10 (3), 367-380.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

Fletcher, J. (1981). The influence of candidates' beliefs and self-presentation strategies in selection interviews. Personnel Review, 10, 14-17.

Fletcher, J., & Spencer, A. (1984). Sex of candidate and sex of interviewer as determinants of self-presentation orientation in interviews: An experimental study. International Review of Applied Psychology, 33, 305-313.

Force has race woes. (1996, September 30). The Vancouver Sun, p. A4.

Frank, F., & Preston, J. (1982). The validity of the assessment center approach and related issues: What are the courts saying? Personnel Administrator.

Frankena, W. 1963. Ethics. New Jersey: Prentice-Hall, Inc.

Friedman, M. (1984). Effect of assessor race and participant race and sex on assessment center ratings. Journal of Assessment Center Technology, 7 (3), 9-14.

Frost, S. (1997). A gender equality analysis. In M. LeBeuf & J. McLean (Eds.), Workshop on women in policing (pp. 68-76). Ottawa: Canadian Police College.

Funston, B., & Meehan, E. (1994). Canada's constitutional law in a nutshell. Ontario: Carswell Legal Publications.

Gale, C. (1983). The predictive validity of an operational assessment center [JIBC Police Academy]. Unpublished master's thesis, University of British Columbia.

Gall, G. (1995). The Canadian legal system (4th ed.). Ontario: Carswell Legal Publications.

Gaugler, B. (1987). Factors affecting assessment center judgments: Rater characteristics and task complexity. Unpublished doctoral dissertation, Colorado State University, Fort Collins.

Gaugler, B., Rosenthal, D., Thornton, G., III, & Bentson, C. (1987). Meta-analysis of assessment center research. Journal of Applied Psychology, 72, 493-511.

Gaugler, B., Thornton, G., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, 74, 611-618.

Gavin, J., & Hamilton, J. (1975). Selecting police using assessment center methodology. Journal of Police Science and Administration, 3 (2), pp. 166-176.

Ghiselli, E., Campbell, J., & Zedeck, S. (1981). Measurement theory for the behavior sciences. San Francisco: W.H. Freeman.

Girodo, M. (1997). Undercover agent assessment centers: Crafting vice and virtue for imposters. Journal of Social Behavior and Personality [Special Edition], 12 (5), 237-260.

Gordon, F. (1997). British Columbia Human Rights Tribunal: Tribunal hearings. In The Continuing Legal Education Society of B.C. Human rights '97 (pp. 3.1.01-05). Vancouver, B.C.: The Continuing Legal Education Society of B.C.

Gottfredson, L. (1988). Reconsidering fairness: A matter of social and ethical priorities. Journal of Vocational Behavior, 33, 293-319.

Green, S., Salkind, N., & Akey, T. (1997). Using SPSS for Windows: Analyzing and understanding data. New Jersey: Prentice Hall, Inc.

Greenwood, J., & McNamara, W. (1967). Interrater reliability in situational tests. Journal of Applied Psychology, 31, 101-106.

Gronlund, N. (1985). Measurement and evaluation in teaching (5th ed.). New York: MacMillan Publishing Co.

Guidelines and ethical considerations for assessment center operations. (1989, May). Seventeenth International Congress on the Assessment Center Method, Pittsburgh, Pennsylvania.

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

Guilford, J., & Fruchter, B. 1978. Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill.

Haas, P., & Springer, J. (1998). Applied policy research: Concepts and cases. New York: Garland Publishing, Inc.

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

Harris, W. (1999). Recruiting women: Are we doing enough? Police: The Law Enforcement Magazine, 23 (8), 18-23

Hinrichs, J. (1978). An eight-year follow-up of a management assessment center. Journal of Applied Psychology, 63 (5), 596-601.

Hinrichs, R., & Haanpera, S. (1976). Reliability of measurement in situational exercises: An assessment of the assessment center method. Personnel Psychology, 29, 31-40.

Hirsh, H., Northrop, L, & Schmidt, F. (1986). Validity generalization results for law enforcement occupations. Personnel Psychology, 39, 399-420.

Hoffman, C., & Thornton, G., III. (1997). Examining selection utility where competing predictors differ in adverse effect. Personnel Psychology, 50 (2), 455-470.

Hogg, P. (1985). Constitutional law of Canada (2nd ed.). Ontario: Carswell Legal Publications.

Holmes, B. (1942). Selection of patrolmen. Journal of Criminal Law, 52.

Hooke, A. (1996). Training police in professional ethics. Journal of Contemporary Criminal Justice, 12 (3), 264-276.

Howard, A. (1974). An assessment of assessment centers. Academy of Management Journal, 17 (1), 115-134.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. Journal of Social Behavior and Personality [Special Edition], 12 (5), 13-52.

Howard, A., & Bray, D. (1988). Managerial lives in transition: Advancing age and changing times. New York: The Guilford Press.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Huck, J. (1973). Assessment centers: A review of external and internal validities. Personnel Psychology, 26, 191-212.

Huck, J. (1974). Determinants of assessment center ratings for white and black females and the relationship of these dimensions to subsequent performance effectiveness. Unpublished doctoral dissertation, Wayne State University, Detroit, Michigan.

Huck, J. (1977). The research base. In J. Moses & W. Byham (Eds.), Applying the assessment center method (pp. 261-291). New York: Pergamon Press.

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

Hurley, K. (1987). Legal aspects of assessment and assessment centers. In H. More & P. Unsinger (Eds.), The police assessment center (pp. 23-48). Illinois: Charles C. Thomas, Publisher.

Hurley, M. (1998). Charter equality rights: Interpretation of section 15 in Supreme Court of Canada decisions (Background Paper No. BP-402E). Ottawa: Library of Parliament, Parliamentary Research Branch.

Hutton, G., & Sampson, F. (1999). The assessment potential and the potential of assessment. Police Chief, 66 (8), 79-83.

Jackson, R. (1939). Reliability of mental tests. British Journal of Psychology, 29, 30-49.

Jaffee, C. (1984). Historical and future perspectives on assessment centers. Journal of Assessment Center Technology, 7 (1), 1-8.

Jaffee, C., Cohen, S., & Cherry, R. (1972, January). Supervisory selection program for disadvantaged or minority employees. Training and Development Journal, 22-27.

Jain, H. (1987). Recruitment of racial minorities in Canadian police forces. Relations Industrielles, 42 (4), 790-804.

James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69, 85-98.

James, L., Demaree, R., & Wolf, G., (1993). r_{wg}: An assessment of within-group interrater agreement. Journal of Applied Psychology, 78 (2), 306-309.

Jensen, A. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. Behavioral and Brain Sciences, 8, 193-219.

Jones, D., & de Villars, A. (1994). Principles of administrative law (2^{nd} ed.). Ontario: Carswell Legal Publications.

Jones, R. (1997). A person perception explanation for validation evidence from assessment centers. Journal of Social Behavior and Personality [Special Edition], 12 (5), 169-178.

Joyce, L., Thayer, P., & Pond, S., III. (1994). Managerial functions: An alternative to traditional assessment center dimensions? Personnel Psychology, 47 (1), 109-121.

Justice Institute of British Columbia Police Academy. (1997). Justice Institute of B.C.: The assessor training program. British Columbia: Author.

Kaplan, R., & Saccuzzo, D. (1982). Psychological testing: Principles, applications, and issues. California: Brooks/Cole Publishing Co.

Kachigan, S. (1986). Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods. New York: Radius Press.

Kachigan, S. (1991). Multivariate statistical analysis: A conceptual introduction (2nd ed.). New York: Radius Press.

Kendall, M. (1948). Rank correlation methods. London: Griffin.

King, M. (1963). Letter from Birmingham Jail [On-line]. Available: www.msstate.edu/Archives/History/USA/Afro-Amer/birmingham.king.

Kleinig, J. (1996). The ethics of policing. New York: Cambridge University Press.

Kleinmann, M., & Koller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. Journal of Social Behavior and Personality [Special Edition], 12 (5), 65-84.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. Personnel Psychology, 40, 243-260.

Klimoski, R., & Strickland, W. (1977). Assessment centers: Valid or merely prescient? Personnel Psychology, 30, 353-361.

Kohlhepp, K. (1992). Assessor accuracy training: A critical component of the assessment center method. Police Chief, 59 (6), 54, 58-60.

Koosis, D. (1997). A self-teaching guide to statistics (4th ed.). Toronto: John Wiley & Sons, Inc.

Kozlowski, S., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. Journal of Applied Psychology, 77 (2), 161-167.

Kuder, G., & Richardson, M. (1937). The theory of estimation of test reliability. Psychometrika, 2, 151-160.

Kudisch, J., Ladd, R., & Dobbins, G. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. Journal of Social Behavior and Personality [Special Edition], 12 (5), 129-144.

Kulis, J. (1987). Issues in assessor selection and training. In H. More & P. Unsinger (Eds.), The police assessment center (pp. 115-144). Illinois: Charles C. Thomas, Publisher.

Landau, S. (Ed.). (1974). Funk & Wagnalls standard college dictionary (Canadian ed.). Toronto: Fitzhenry & Whiteside Ltd.

Landrine, H., & Klonoff, E. (1997). Discrimination against women: Prevalence, consequences, remedies. California: Sage Publication, Inc.

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

Lawlis, F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. Psychological Bulletin, 78, 17-20.

LeBeuf, M., & McLean, J. (Eds.). (1997). Women in policing in Canada: Beyond the year 2000—its challenges. Workshop on Women in Policing. Ottawa: Canadian Police College.

Lee, C. (1985). Assessment centers: A method with proven mettle. Training, 22 (7), 69-70.

Lee, P. (1997). Bayesian statistics: An introduction (2^{nd} ed.). NewYork: Oxford University Press.

Lerner, M. (1982). The justice motive in human relationships and the economic model of man: A radical analysis of fact and fictions. In V. Derlega & J. Grezlak (Eds.), Cooperation and helping behavior: Theory and research (pp. 121-145). New York: Academic Press.

Leventhal, G., Karuza, J., & Fry, W. (1980). Beyond fairness: A theory of allocation preferences. In G. Mikula (Ed.), Justice and social interaction (pp. 167-218). New York: Springer-Verlag.

Li, H., Rosenthal, R., & Rubin, D. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. Psychological Methods, 1 (1), 98-107.

Lind, E., & Tyler, T. (1988). The social psychology of procedural justice. New York: Plenum Press.

Linden, R. (1983). Women in policing: A study of lower mainland R.C.M.P. detachments. Canadian Police College Journal, 3.

Linden, R., & Minch, C. (1994). Women in policing: A review: 1984-92 (Programs Branch User Report). Ottawa: Solicitor General Canada.

Loacher, G. (1974). Assessment procedures used as educational criterion. Assessment & Development, 2 (1), 5.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum.

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lovett, D. (1997). Human rights in the workplace: The duty to accommodate. In The Continuing Legal Education Society of B.C. Human rights '97 (pp. 2.1.01-29). Vancouver, B.C.: The Continuing Legal Education Society of B.C.

Lowry, P. (1996, Fall). A survey of the assessment center process in the public sector. Public Personnel Management, 25 (3), 307-321.

Lowry, P. (1993, Fall). The assessment center: An examination of the effects of assessor characteristics on assessor scores. Public Personnel Management, 22 (3), 487-501.

Lowry, P. (1994, Fall). Selection methods: comparison of assessment centers with personnel records evaluations. Public Personnel Management, 23 (3), 383-395.

Lowry, P. (1997). The assessment center process: New directions. Journal of Social Behavior and Personality [Special Edition], 12 (5), 53-62.

Lu, K. (1971). A measure of agreement among subjective judgments. Educational and Psychological Measurement, 31, 75-84.

MacKinnon, D. (1975a, May). An overview of assessment centers. Center for Creative Leadership Technical Report No. 1. Adapted from a paper presented at the Industrial Psychologists Meeting, Center for Creative Leadership, Greensboro, North Carolina, January 29-30, 1975.

MacKinnon, D. (1975b). An overview of assessment centers. University of California, Berkeley: Center for Creative Leadership.

MacKinnon, D. (1977). From selecting spies to selecting managers—The OSS assessment program. In J. Moses & W. Byham (Eds.), Applying the assessment center method (pp. 13-30). New York: Pergamon Press.

McClellan, E. (1985). The assessment center method: Its development and use. British Columbia: Justice Institute of British Columbia Police Academy.

McConnell, J., & Parker, T. (1972). An assessment center program for multiorganizational use. Training and Development Journal, 26 (3), 6-14.

McGhee, A., & Deen, M. (1979). Utilizing the assessment center to select police officers for the Ocala, Florida Department. Police Chief, 46 (8), 69-74.

McGinnis, J. (1987). Validity in the assessment center. In H. More & P. Unsinger (Eds.), The police assessment center (pp. 91-113). Illinois: Charles C. Thomas, Publisher.

McGinnis, J., & Carpenter, G. (1980). The Canadian Police College pilot municipal [police] force assessment center. Canadian Police College Journal, 4, (1), 1-31.

McKelvie, S. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. Psychological Reports, 65, 161-162.

McLean, J. (1997). The future of women in policing in Canada. In M. LeBeuf & J. McLean (Eds.), Workshop on women in policing (pp. 176-182). Ottawa: Canadian Police College.

Magaldi, R., Mendoza, R., Jr., Stafford, G., & Frank, R. (1984). Police promotional level assessment centers: The Metro-Date police department experience—focus on race, sex, and the assessment center cycle. Journal of Assessment Center Technology, 7 (2), 9-16.

Maher, P. (1984, May). Assessing assessment centers: An introduction. American Fire Journal, pp. 20-25.

Martin, S. (1989). Female officers on the move? A status report of women in policing. In R. Dunham & G. Alpert (Eds.), Critical issues in policing. Prospect Heights, Illinois: Waveland.

Martin, S. (1991). The effectiveness of affirmative action: The case of women in policing. Justice Quarterly, 8 (4), 489-504.

Martin, P., & Bateson, R. (1993). Measuring behavior: An introductory guide (2nd ed.). Cambridge: Cambridge University Press.

Mayes, B. (1997). Insights into the history and future of assessment centers: An interview with Dr. Douglas W. Bray and Dr. William Byham. Journal of Social Behavior and Personality [Special Edition], 12 (5), 3-12.

Mehrens, W., & Lehmann, I. (1978). Measurement and evaluation in education and psychology (2nd ed.). New York: Holt, Rinehart and Winston.

Mendenhall, M. (1989, September). Successful legal defense of the assessment center method: A first hand account. Paper presented at the International Conference on Assessment Centers, Miami, Florida.

Mendenhall, M. (1992, June). Successful legal defense of the assessment center method. The Police Chief, pp. 61-63.

Mewett, A. (1999). Secondary fact, prejudice and stereotyping. Criminal Law Quarterly, 42 (2 & 3), 319-337.

Miller, S. (1998). Rocking the rank and file: Gender issues and community policing. Journal of Contemporary Criminal Justice, 14 (2), 156-172.

Miller, L., & Whitehead, J. 1996. Introduction to criminal justice research and statistics. Ohio: Anderson Publishing Co.

Mills, R. (1976). Simulated stress in police recruit selection. Journal of Police Science and Administration, 4, 179-186.

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

Mood, M. (1983). Introduction to policy analysis. New York: Holland.

Moore, D. (1985). Statistics: Concepts and controversies (2nd ed.). New York: W.H. Feeman and Company.

Moore, H. (1997). An historical account of women in policing in Canada. In M. LeBeuf & J. McLean (Eds.), Workshop on women in policing (pp. 36-45). Ottawa: Canadian Police College.

Moses, J. (1971). Assessment center performance and management progress (ATT). Paper presented at symposium, "Validity of Assessment Centers," at the 79th Annual Convention of the American Psychological Association.

Moses, J. (1972). Assessment center performance and management progress. Studies in Personnel Psychology, 4 (1), 7-12.

Moses, J. (1973). The development of an assessment center for the early identification of supervisory potential. Personnel Psychology, 26, 569-580.

Moses, J. (1977). The assessment center method. In J. Moses & W. Byham (Eds.), Applying the assessment center method (pp. 3-11). New York: Pergamon Press.

Moses, J., & Boehm, V. (1975). Relationship of assessment-center performance to management progress of women. Journal of Applied Psychology, 60 (4), 527-529.

Moses, J., & Byham, W. (Eds.). (1977). Applying the assessment center method. New York: Pergamon Press.

Murray, H. (1938). Explorations in personality. New York: Oxford University Press.

Murray, J. (1996). Gender effects in assessment center settings. Paper presented at the International Congress on the Assessment Center Method, Washington, D.C.

Nelson, E. (1992). Employment equity and the Red Queen's hypothesis: Recruitment and hiring in western Canadian municipal police departments. Canadian Police College Journal, 16 (3), 184-203.

Norton, S. (1981). The assessment center process and content validity: A reply to Dreher and Sackett. Academy of Management Review, 6 (4), 561-566.

Novick, M., & Lewis, G. (1967). Coefficient alpha and the reliability of composite measures. Psychometrika, 32, 1-13.

Nunnally, J. (1978). Psychometric theory. New York: McGraw-Hill.

O'Hara, K., & Love, K. (1987). Accurate selection of police officials within small municipalities: "Et tu assessment center?" Public Personnel Management, 16 (1), 9-14.

O'Leary, L. (1997, May). The current application of assessment centers by police departments in the United States. Paper presented at the International Congress on the Assessment Center Method, London, England.

Office of Strategic Services (OSS) Assessment Staff. (1948). Assessment of men. New York: Rinehart.

Overall, J. (1965). Reliability of composite ratings. Educational and Psychological Measurement, 25, 1011-1022.

Paetzold, R., & Willborn, S. (1999). The statistics of discrimination: Using statistical evidence in discrimination cases. MN: West Group.

Parker, K., Hanson, R., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. Psychological Bulletin, 103, 367-373.

Parker, T. (1980, February). Assessment centres: A statistical study. The Personnel Administrator, 65-67.

Parkinson, C. (1957). Parkinson's Law. New York: Ballantine.
Pergamon Press.

Peters, L., O'Conner, E., Weekley, J., Pooyan, A., Frank, B., & Erenkrantz, B. (1984). Sex bias and managerial evaluations: A replication and extension. Journal of Applied Psychology, 69, 349-352.

Policing in British Columbia Commission of Inquiry (1994). Closing the gap: Policing and the community. The report. Province of British Columbia.

Polisar, J., & Milgram, D. (1998). Recruiting, integrating and retaining women police officers: Strategies that work. Police Chief, 65 (10), 42-52.

Polowek, K. (1996). Retention of British Columbia's municipal police officers: An examination of reasons for leaving. British Columbia: Police Commission.

Prindiville, J. (1975). Women in policing in British Columbia. British Columbia: Police Commission.

Proposed guidelines for recruitment and selection of visible minority police officers in Canada. Currents, 3 (4), 13-16.

Pynes, J. (1988). The predictive validity of an assessment center to select entry-level law enforcement officers. Unpublished doctoral dissertation, Florida Atlantic University, Boca Raton.

Pynes, J., & Bernardin, H. (1989). Predictive validity of an entry-level police officer assessment center. Journal of Applied Psychology, 74 (5), 831-833.

Pynes, J., & Bernardin, H. (1992). Entry-level police selection: The assessment center as an alternative. Journal of Criminal Justice, 20 (1), 41-52.

Quarles, C. (1982). Assessment center as a managerial success predictor. Journal of Security Administration, 5 (2), 81-87.

Rachels, J. (1999). The elements of moral philosophy (3rd ed.). New York: McGraw-Hill College.

Rajaratnam, N. (1960). Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 25, 261-271.

Rank objections: Study calls growth of women in policing "alarmingly slow." (1998, May 15). Law Enforcement News, p. 1.

Rawls, J. (1993). Political liberalism. New York: Columbia University Press.

Ree, M., & Earles, J. (1992). Intelligence is the best predictor of job performance. Current Directions in Psychological Science, 1 (3), 86-89.

Reilly, R., Henry, S., & Smither, J. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. Personnel Psychology, 43, 71-84.

Ritchie, R., & Moses, J. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. Journal of Applied Psychology, 68 (2), 227-231.

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centers: Dimensions into exercises won't go. Journal of Occupational Psychology, 60, 187-195.

Rockowitz, M., Shuttleworth, D., Shukyn, M., Brownstein, & Peters, M. (1998). How to prepare for the GED high school equivalency exam (3rd ed. (Canadian)). Hauppauge, New York: Barron's Educational Series, Inc.

Rosenthal, R. (1987). Judgment studies: Design, analysis, and meta-analysis. Cambridge: Cambridge University Press.

Rosenthal, R., & Rosnow, R. (1991). Essentials of behavior research: Methods and data analysis (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. (1982). A simple general purpose display of magnitude of experimental effect. Journal Of Educational Psychology, 74, 166-169.

Rosnow, R., & Rosenthal, R. (1996). Beginning behavioral research: A conceptual primer (2nd ed.). New Jersey: Prentice Hall.

Ross, J. (1980). Determination of the predictive validity of the assessment center approach to selecting police managers. Journal of Criminal Justice, 8 (2), 89-96.

SPSS® Base 10.0 Applications Guide. (1999). Chicago, Illinois: SPCC Inc.

Sackett, P., & Dreher, G. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67 (4), 401-410.

Sackett, P., & Hakel, M. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. Organizational Behavior and Human Performances, 23, 120-137.

Schmidt, F., Berner, J., & Hunter, J. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58.

Schmidt, F., & Hunter, J. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. Journal of Applied Psychology, 74, 368-370.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. Journal of Applied Psychology, 63, 171-176.

Schmitt, N., Ford, J., & Stultz, D. (1986). Changes in self-perceived ability as a function of performance in an assessment center. Journal of Occupational Psychology, 59, 327-335.

Schmitt, N., Gooding, R., Noe, R., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37 (3), 407-422.

Seagram, B., & Stark-Adamec, C. (1992). Women in Canadian policing: Why are they leaving? Police Chief, 59, 120-128.

Segrave, K. (1995). Policewomen: A history. North Carolina: McFarland & Company, Inc.

Shechtman, Z. (1992). A group assessment procedure as a predictor of on-the-job performance of teachers. Journal of Applied Psychology, 77 (3), 383-387.

Sheppard, C. (1993). Litigating the relationship between equity and equality (Study Paper). Ontario: Ontario Law Reform Commission.

Shore, T., & Tashchian, A., & Adams, J. (1997). The role of gender in a developmental assessment center. Journal of Social Behavior and Personality [Special Edition], 12 (5), 191-203.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86 (2), 420-428.

Sidgwick, H. (1907). The methods of ethics (7th ed.). London: Macmillan and Company, Ltd.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw Hill.

Silzer, R. (1985). Assessment center validity across two organizations. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, California.

Singer, M. (1993). Fairness in personnel selection: An organizational justice perspective. Brookfield, USA: Avebury.

Snedecor, G., & Cochran, W. (1980). Statistical methods (7th ed.). Iowa: Iowa State University Press.

Soltan, A. (1994). Human rights. In The Continuing Legal Education Society of B.C. Employment law—1994 (pp. 4.1.01-4.1.32). Vancouver, B.C.: The Continuing Legal Education Society of B.C.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 171-195.

Spychalski, A., Quinones, M., Gaugler, B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. Personnel Psychology, 50 (1), 71-90.

Standards and ethical considerations for assessment center operations. (1975, May). Third International Congress on the Assessment Center Method, Quebec, Canada.

Standards and ethical considerations for assessment center operations. (1979, June). Seventh International Congress on the Assessment Center Method, New Orleans, Louisiana.

Statistics Canada. (1999). Police resources in Canada, 1999. Ottawa: Canadian Centre for Justice Statistics.

Status of Women Canada. (1996). Gender-based analysis: A guide for policy-making (Working Document). Ottawa: Author.

Stinchcomb, J. (1985, June). Why not the best? Using assessment centers for officer selection. Corrections Today, pp. 122-124.

Tanovich, D. (1999). Annual review of criminal law: 1998. Ontario: Carswell Legal Publications.

Taylor, K. (1983). The selection and training of police officers in British Columbia. Police Studies, 6 (3), 44-49.

Teel, K., & DuBois, H. (1983). Participants' reactions to assessment centers. Personnel Administrator (March), 85-91.

Tenopyr, M. (1977). Content-construct confusion. Personnel Psychology, 30, 47-54.

Thibaut, J. & Walker, L. (1978). A theory of procedure. California Law Review, 66, 541-566.

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

Thornton, C., III, Tziner, A., Dahan, M., Clevenger, J., & Meir, E. (1997). Construct validity of assessment center judgments: Analyses of the behavioral reporting method. Journal of Social Behavior and Personality [Special Edition], 12 (5), 109-128.

Tielsch, G., & Whisenand, P. (1977). The assessment center approach in the selection of police personnel. California: Davis Publishing Company, Inc.

Tinsley, H., & Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22 (4), 358-376.

Traub, R. (1994). Reliability for the social sciences: Theory and applications. California: Sage Publications, Inc.

Turner, T. (1978, Autumn). The assessment center program: A new way of matching the right person to the right job. British Columbia Police Journal, pp. 12-16.

Turner, T., & Higgins, K. (1977, July). Initial selection assessment center program: Police constable job analysis report. British Columbia: Police Commission.

Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. Journal of Applied Psychology, 67, 728-736.

Vasan, R. (Ed.). (1980). The Canadian law dictionary. Ontario: Law & Business Publications (Canada), Inc.

Verdun-Jones, S. (1997). Criminal law in Canada: Cases questions, and the Code (2nd ed.). Canada: Harcout Brace.

Vining, A., McPhillips, D., & Boardman, A. (1986). Use of statistical evidence in employment discrimination litigation. The Canadian Bar Review, 64, 660-702.

Vizkelety, B. (1987). Proving discrimination in Canada. Ontario: Carswell Legal Publications.

Vito, G., & Latessa, E. (1989). Statistical applications in criminal justice. California: Sage Publications.

Vlastos, G. (1970). Justice and equality. In A. Melden (Ed.), Human rights (pp. 76-95). California: Wadsworth Publishing Company, Inc.

Waddams, S. (1997). Introduction to the study of law (5th ed.). Ontario: Carswell Legal Publications.

Wagner, J. (1993). Ignorance in educational research: Or, how can you not know that? Educational Researcher, 22 (5), 15-23.

Walker, G. (1993). The status of women in Canadian policing: 1993 (No. 1993-22). Ottawa: Solicitor General Canada.

Walker, S. (1993). Taming the systems: The control of discretion in criminal justice 1950-1990. New York: Oxford University Press.

Walsh, J., Weinberg, R., & Fairfield, M. (1987). The effects of gender on assessment center evaluations. Journal of Occupational Psychology, 60 (4), 305-309.

Wasserstrom, R. (1970). Rights, human rights, and racial discrimination. In A. Melden (Ed.), Human rights (pp. 96-110). California: Wadsworth Publishing Company, Inc.

Watson, P., & Barber, B. (1988). The struggle for democracy. Toronto: Lester & Orpen Dennys Ltd.

Weiner, N. (1993). Employment equity: Making it work. Canada: Butterworths.

Weissbrodt, D. (1988). Human rights: An historical perspective. In P. Davies (Ed.), Human rights (pp. 1-20). London: Routledge.

Whitehead, A. (1933). Adventures of ideas. Toronto: Macmillan Co.

Whitehurst, G. (1985). On lies, damned lies, and statistics: Measuring interrater agreement. American Psychologist, 40 (5), 568-569.

Wigdor, A., & Garner, W. (1982). Ability testing: Uses, consequences and controversies. Washington, DC: Academy Press.

Wildavsky, A. (1979). Speaking truth to power: The art and craft of policy analysis. Boston: Little, Brown and Co.

Wildt, A., & Ahtola, O. (1978). Analysis of covariance (series no. 07-012). California: Sage.

Winer, B. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Wollowick, H., & McNamara, W. (1969). Relationships of the components of an assessment center to management success. Journal of Applied Psychology, 53, 348-352.

Women climbing police ranks slowly. (1998, August 28). The Globe and Mail, p. A2.

Women in policing. (1998). Police Chief, 65 (10), 36-40.

Women still penalized in promotions, study shows. (1997, December 17). The Vancouver Sun, p. A1.

Wonderlic personnel test & scholastic exam: User's manual. 1992. Illinois: Wonderlic Personnel Test, Inc.

Yan, T., & Slivinski, L. (1976). A history of the assessment centre method in the military. Ottawa: Queen's Printer.

Yogis, R. (1983). Canadian law dictionary. Toronto: Barron's Educational Series, Inc.

Zeisel, H., & Kaye, D. (1997). Prove it with figures: Empirical methods in law and litigation. New York: Springer-Verlag, Inc.

Zinn, R., & Brethour, P. (1996). The law of human rights in Canada: Practice and procedure. Ontario: Canada Law Book.

APPENDIX 1

Total Cases

	Total Cases:[a] 1978-1999[b]					
	Included		Excluded		Total	
	N	%	N	%	N	%
Age						
Male	952	32.21	1526	51.62	2478	83.83
Female	214	7.24	259	8.76	473	16.00
Missing	n/a	n/a	5	0.17	5	0.17
Total	1166	39.45	1790	60.55	2956	100
Education						
Male	898	30.37	1580	53.45	2478	83.83
Female	206	6.97	267	9.03	473	16.00
Missing	n/a	n/a	5	0.17	5	0.17
Total	1104	37.34	1852	62.65	2956	100
Exam						
Male	171	5.79	2307	78.05	2478	83.83
Female	48	1.62	425	14.37	473	16.00
Missing	n/a	n/a	5	0.17	5	0.17
Total	219	7.41	2737	92.59	2956	100

Notes:
[a] Minor discrepancies in adding are the result of rounding at two places after the decimal point. Also as a result of rounding, the missing data may appear as .17 or .20 in subsequent tables.
[b] Up to October, 1999.

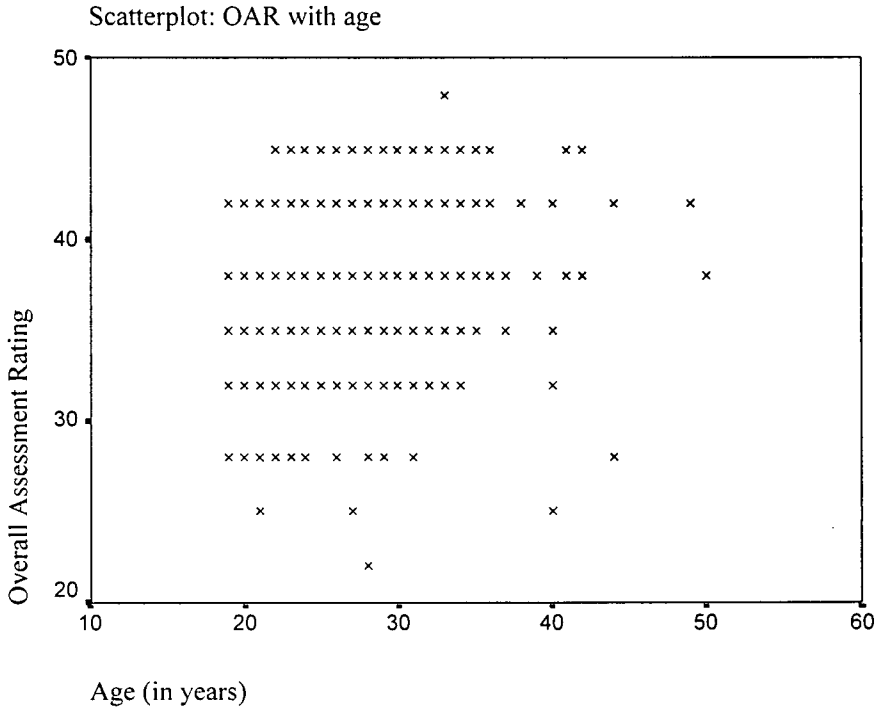
APPENDIX 2

Sample Summary

		Age	Education	Exam
Male	Mean	25.61	2.50	70.19
	N	952	898	171
	Sd	4.03	1.65	8.55
	Minimum	19	0	46
	Maximum	50	8	93
Female	Mean	26.49	3.24	73.48
	N	214	206	48
	Sd	3.82	1.53	7.75
	Minimum	20	0	60
	Maximum	42	5	92
Total	Mean	25.77	2.64	70.91
	N	1166	1104	219
	Sd	4.00	1.66	8.47
	Minimum	19	0	46
	Maximum	50	8	93

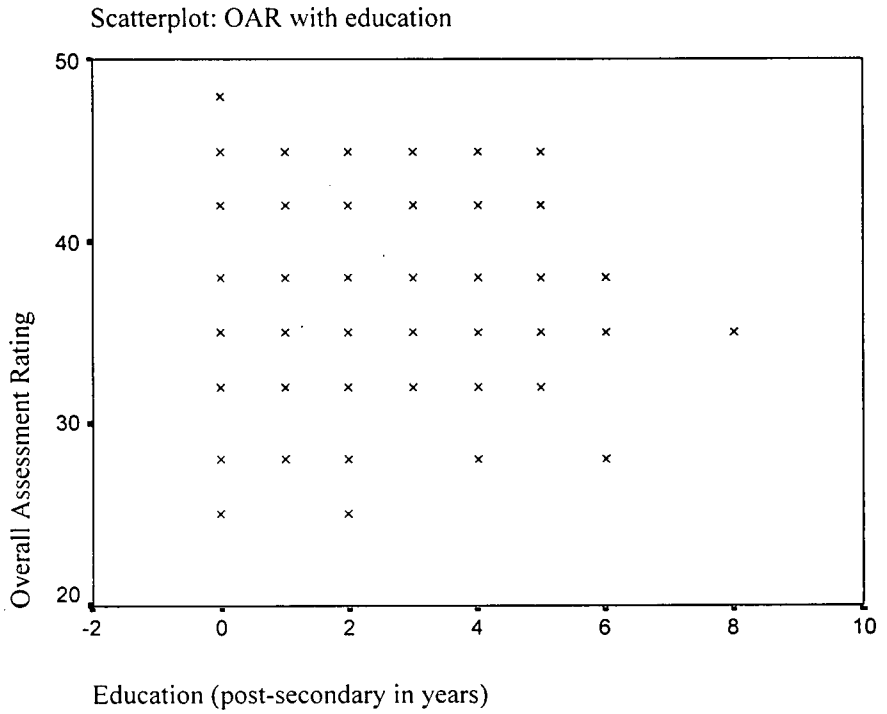
APPENDIX 3

Scatterplot: OAR with Age



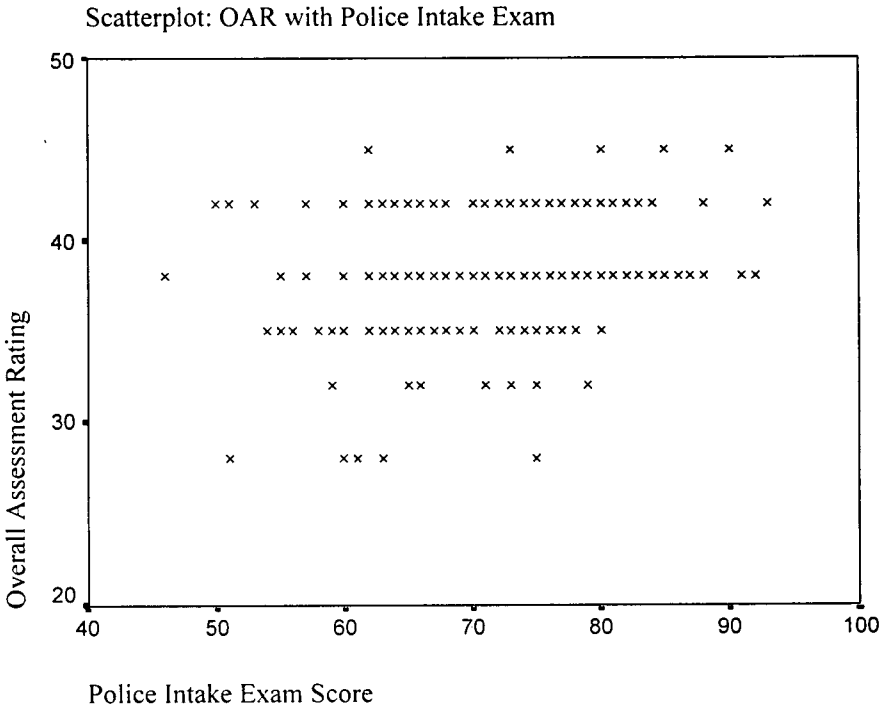
APPENDIX 3 cont.

Scatterplot: OAR with Education



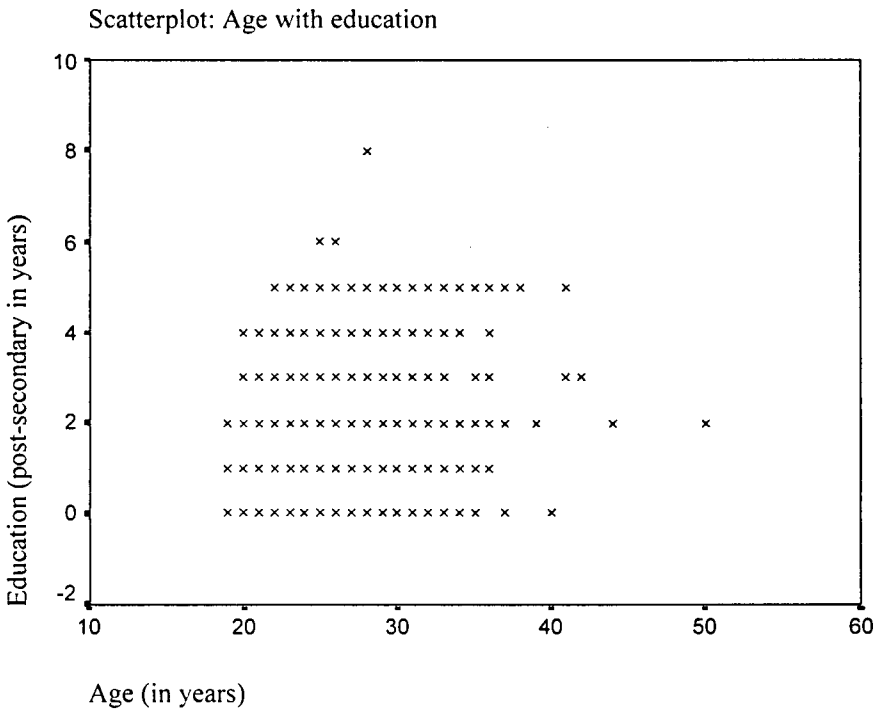
APPENDIX 3 cont.

Scatterplot: OAR with Police Intake Exam



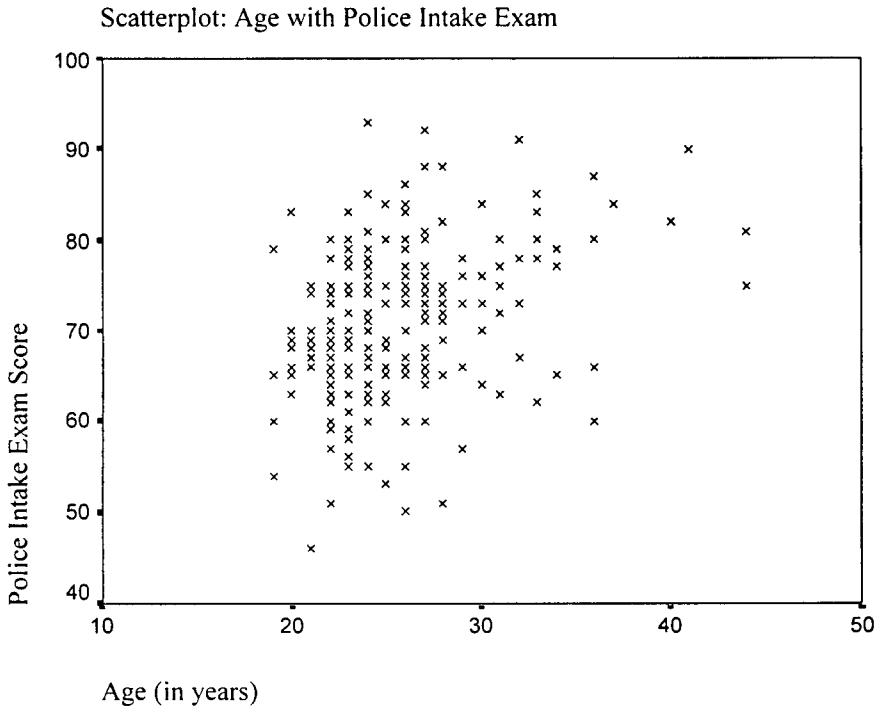
APPENDIX 3 cont.

Scatterplot: Age with Education



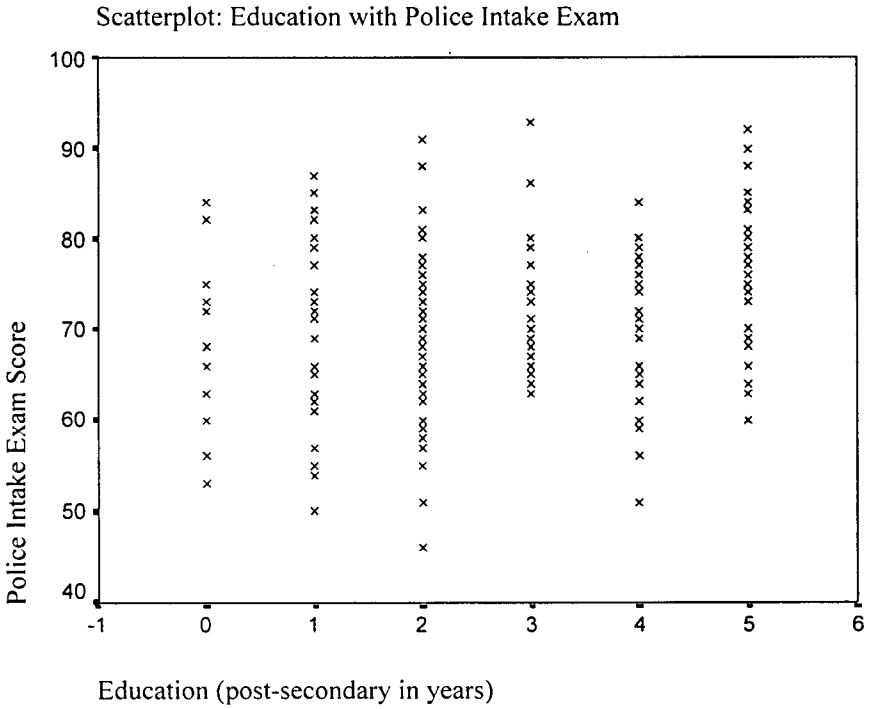
APPENDIX 3 cont.

Scatterplot: Age with Police Intake Exam



APPENDIX 3 cont.

Scatterplot: Education with Police Intake Exam



APPENDIX 4

Correlation Matrix: Dimensions with OAR

Part A

	1 Able	2 Deci	3 Fact	4 Flex	5 Init	6 Integ	7 Inter	8 Matu
1 Able Learn	1.00	.461	.580	.085	.475	.234	.264	.418
p (2-tail)[b]		.000	.000	.148	.000	.000	.000	.000
N	290	290	290	290	290	290	290	96
2 Decisiveness		1.00	.522	-.073	.599	.214	.283	.451
p (2-tail)			.000	.213	.000	.000	.000	.000
N		290	290	290	290	290	290	96
3 Fact Finding			1.00	.083	.517	.272	.262	.344
p (2-tail)				.160	.000	.000	.000	.001
N			290	.290	290	290	290	96
4 Flexibility				1.00	.069	.001	.279	.367
p (2-tail)					.244	.983	000	.000
N				290	290	290	290	96
5 Initiative					1.00	.313	.356	.504
p (2-tail)						.000	.000	.000
N					290	290	290	96
6 Integrity						1.00	.153	.248
p (2-tail)							.009	.015
N						290	290	96
7 Interpersonal							1.00	.691
p (2-tail)								.000
N							290	96
8 Maturity								1.00
p (2-tail)								
N								96

Notes:

a Pairwise.

[b] With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0004 (.05/136).

APPENDIX 4 cont.

Part B

	9 Oral	10 Pers	11 Prac	12 Prob	13 Stres	14 Writ	15[c] Toler	16[c] Auth	17 OAR
1 Able to learn	.463	.489	.628	.527	.517	.236	.330	.029	.589
p (2-tail)[b]	.000	.000	.000	.000	.000	.000	.000	.692	.000
N	290	289	290	290	290	290	290	194	290
2 Decisiveness		.620	.530	.688	.638	.147	.341	.124	.658
p (2-tail)		.000	.000	.000	.000	.012	.000	.084	.000
N		289	290	290	290	290	290	194	290
3 Fact finding			.555	.561	.541	.178	.312	.078	.595
p (2-tail)			.000	.000	.000	.002	.000	.280	.000
N			290	290	290	290	290	194	290
4 Flexibility				.035	.102	-.015	.264	.042	.141
p (2-tail)				.548	.084	.802	.000	.563	.017
N				290	290	290	290	194	290
5 Initiative					.599	.232	.387	.182	.716
p (2-tail)					.000	.000	.000	.011	.000
N					290	290	290	194	290
6 Integrity						.168	.245	.462	.490
p (2-tail)						.004	.000	.000	.000
N						290	290	194	290
7 Interpersonal							.730	.083	.526
p (2-tail)							.000	.251	.000
N							290	194	290
8 Maturity									.772
p (2-tail)									.000
N								0	96

Notes:
a Pairwise.
[b] With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0004 (.05/136).
[c] In 1994, interpersonal tolerance and interpersonal sensitivity were collapsed into one variable (i.e., interpersonal sensitivity), and adherence to authority was replaced with maturity.

APPENDIX 4 cont.

Part C

	1 Able	2 Deci	3 Fact	4 Flex	5 Init	6 Integ	7 Inter	8 Matu
9 Oral Comm		.628	.555	.100	.675	.283	.463	.668
p (2-tail)[b]		.000	.000	.089	.000	.000	.000	.000
N		290	290	290	290	290	290	96
10 Personal Im			.556	.055	.660	.411	.439	.708
p (2-tail)			.000	.351	.000	.000	.000	.000
N			289	289	289	289	289	96
11 Practical Int				.080	.551	.371	.393	.694
p (2-tail)				.174	.000	.000	.000	.000
N				290	290	290	290	96
12 Probl'm Con					.647	.356	.395	.510
p (2-tail)					.000	.000	.000	.000
N					290	290	290	96
13 Stress Toler						.392	.436	.602
p (2-tail)						.000	.000	.000
N						290	290	96
14 Writt'n Com							.074	.053
p (2-tail)							.208	.611
N							290	96
15 Tolerance[c]								.691
p (2-tail)								.000
N								96
16 Authority[c]								
p (2-tail)								
N								
17 OAR								
p (2-tail)								
N								

Notes:

a Pairwise.

[b] With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0004 (.05/136).

[c] In 1994, interpersonal tolerance and interpersonal sensitivity were collapsed into one variable (i.e., interpersonal sensitivity), and adherence to authority was replaced with maturity.

APPENDIX 4 cont.

Part D

	9 Oral	10 Pers	11 Prac	12 Prob	13 Stres	14 Writ	15[c] Toler	16[c] Auth	17 OAR
9 Oral Comm	1.00	.783	.603	.651	.669	.219	.475	.150	.719
p (2-tail)[b]		.000	.000	.000	.000	.000	.000	.037	.000
N	290	289	290	290	290	290	290	194	290
10 Personal Im		1.00	.616	.687	.703	.203	.459	.248	.792
p (2-tail)			.000	.000	.000	.001	.000	.001	.000
N		289	289	289	289	289	289	193	289
11 Practical Int			1.00	.619	.583	.266	.480	.189	.719
p (2-tail)				.000	.000	.000	.000	.008	.000
N			290	290	290	290	290	194	290
12 Probl'm Con				1.00	.699	.223	.456	.214	.756
p (2-tail)					.000	.000	.000	.003	.000
N				290	290	290	290	194	290
13 Stress Toler					1.00	.214	.479	.138	.745
p (2-tail)						.000	.000	.055	.000
N					290	290	290	194	290
14 Writt'n Com						1.00	.142	.068	.278
p (2-tail)							.015	.345	.000
N						290	290	194	290
15 Tolerance[c]							1.00	.247	.561
p (2-tail)								.001	.000
N							290	194	290
16 Authority[c]								1.00	.284
p (2-tail)									.000
N								194	194
17 OAR									1.00
p (2-tail)									
N									290

Notes:
a Pairwise.
[b] With a Bonferroni adjustment (significance level divided by total number of correlations), the significance level for .05 would be .0004 (.05/136).
[c] In 1994, interpersonal tolerance and interpersonal sensitivity were collapsed into one variable (i.e., interpersonal sensitivity), and adherence to authority was replaced with maturity.